APPLICATION OF OPTIMIZATION ALGORITHMS IN CHEMISTRY

EDITED BY: Jorge M. C. Marques, Emilio Martinez-Nunez and William L. Hase

PUBLISHED IN: Frontiers in Chemistry

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88963-710-2
DOI 10.3389/978-2-88963-710-2

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

APPLICATION OF OPTIMIZATION ALGORITHMS IN CHEMISTRY

Topic Editors:
Jorge M. C. Marques, University of Coimbra, Portugal
Emilio Martinez-Nunez, Universidade de Santiago de Compostela, Spain
William L. Hase, Texas Tech University, United States

This eBook is dedicated to Prof. William L. Hase, who passed away on Monday, March 23, 2020.

Citation: Marques, J. M. C., Martinez-Nunez, E., Hase, W. L., eds. (2020). Application of Optimization Algorithms in Chemistry. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88963-710-2

Table of Contents

frontiers in Chemistry

EDITORIAL
published: 20 March 2020
doi: 10.3389/fchem.2020.00198

Editorial: Application of Optimization Algorithms in Chemistry

Jorge M. C. Marques[1]*, *Emilio Martínez-Núñez*[2] *and William L. Hase*[3]

[1] Coimbra Chemistry Centre (CQC), Department of Chemistry, University of Coimbra, Coimbra, Portugal, [2] Departmento de Química Física, Facultade de Química, Universidade de Santiago de Compostela, Santiago de Compostela, Spain, [3] Department of Chemistry and Biochemistry, Texas Tech University, Lubbock, TX, United States

Keywords: molecular geometry search, fitting potential energy functions, spectral assignment, global optimization algorithms, energy landscapes

Editorial on the Research Topic

Application of Optimization Algorithms in Chemistry

Molecular structure optimization, fitting potential energy functions to *ab initio* and experimental data, and spectral assignment are among the hardest optimization tasks in molecular sciences. These are fundamental problems in chemistry, but they can also be relevant in molecular physics and biochemistry. In past decades, several methodologies have been proposed to help in the above mentioned tasks, and some of them are already incorporated into computational tools, such as GMIN (Wales and Scheraga, 1999; Wales, 2010), Gradient Embedded Genetic Algorithm or GEGA (Alexandrova and Boldyrev, 2005), OGOLEM (Hartke, 1993; Dieterich and Hartke, 2017), Birmingham Cluster Genetic Algorithm or BCGA (Johnston, 2003; Shayeghi et al., 2015), Evolutionary Algorithm for Molecular Clusters or EA_MOL (Llanio-Trujillo et al., 2011; Marques and Pereira, 2011), Global Reaction Route Mapping or GRRM (Ohno and Maeda, 2006, 2019), Automated Mechanisms and Kinetics or AutoMeKin (Martínez-Núñez, 2015a,b; Martínez-Núñez, 2020), and Genetic Algorithm fitting or GAFit (Rodríguez-Fernández et al., 2017, 2020). Most of these computational programs are interfaced with well-known packages that perform electronic-structure calculations and, hence, allow for a direct assessment of the semi-empirical, density functional theory (DFT) or *ab initio* energy of the system during the optimization process. Another relevant methodology to explore low-energy landscapes is the parallel-tempering Monte Carlo technique, which has been also applied in the calculation of thermodynamic properties.

Global geometry optimization studies are, now, being extended to systems of increasing complexity. In particular, global optimization algorithms have been applied to a great diversity of chemical systems, including atomic and molecular clusters as well as colloidal aggregates and biomolecules. Nonetheless, optimization work needs, in general, a large number of computational resources and, hence, improvements in algorithms to relieve the burden. Major challenges are concerned with the treatment of systems with increasing size and incorporating higher levels of theory in the molecular model. Also, multi-component aggregates pose an important combinatorial problem and require novel optimization strategies. Although the use of state-of-the-art spectroscopic techniques to probe the structure of clusters has allowed for close collaborative work involving computational and experimental achievements, there is still room for greater improvement in this effort. In particular, comparisons between theoretical and experimental spectroscopic data will benefit from significant improvements in algorithms devoted for the spectral assignment.

OPEN ACCESS

Edited and reviewed by:
Sam P. De Visser,
University of Manchester,
United Kingdom

****Correspondence:***
Jorge M. C. Marques
qtmarque@ci.uc.pt

Specialty section:
This article was submitted to Theoretical and Computational Chemistry,
a section of the journal Frontiers in Chemistry

Received: *22 February 2020*
Accepted: *04 March 2020*
Published: *20 March 2020*

Citation:
Marques JMC, Martínez-Núñez E and Hase WL (2020) Editorial: Application of Optimization Algorithms in Chemistry. Front. Chem. 8:198. doi: 10.3389/fchem.2020.00198

Pursuing those purposes, we believe the collection of papers for the present Research Topic illustrates the broad scope of computational strategies for global optimization applications in chemistry. All contributions describe optimization strategies for a great diversity of chemical systems.

Basin-hopping (BH) is able to generate a coarse-grained mapping of a potential energy surface (PES) in terms of local minima, which can then be used to gain insights into molecular dynamics and thermodynamic properties as pointed out by Zhou et al. in their contribution. These authors also show how unsupervised machine learning tools can be employed to enhance BH searches, which result in more efficient identification of local minima and transition states connecting them.

Jana et al. employ a particle swarm optimization (PSO) method to search for small C_n clusters. PSO is another useful algorithm for a stochastic search in multidimensional space. The method has proven efficient in hard optimization problems compared with traditional methods.

Hernández-Rojas and Calvo also employ BH method, this time to predict low-energy structures of adamantane clusters by using both coarse-grained and atomistic potential models. Although coarse-grained models are appealing for the complex clusters that are studied, the comparison with atomistic potentials shows that some relevant structural details are not captured by the former.

As for seeking conformational minima of flexible acyclic molecules, Ferro-Costas and Fernández-Ramos propose an algorithm that combines a systematic variation of torsion angles with a Monte Carlo search. This methodology has been applied to calculate multi-structural partition functions of several alcohols ranging from n-propanol to n-heptanol and was also tested with the amino acid L-serine.

Panadés-Barrueta et al. put forward a fully automated method to generate highly-accurate semiempirical potential energy surfaces. They use global optimization techniques and automated PES sampling algorithms to refine specific reaction parameters of semi-empirical Hamiltonians, which can be subsequently employed in quantum dynamics studies.

In turn, Wang et al. carry out a microsolvation study of Na^+ with water by applying a genetic algorithm combined with density functional theory to obtain low-energy structures of the clusters. Also, a new genetic algorithm is proposed by Silva et al. for the prediction of structures of nanoparticles. This work explores the efficacy of new evolutionary operators to treat Lennard-Jones and carbon clusters.

Khatun et al. develop a global optimizer which grows the cluster by adding atoms one by one. The method is tested by studying transition-metal clusters and binary and ternary nanoalloys of such elements.

Cova and Pais review and discuss deep learning strategies for optimizing the prediction of chemical patterns, which includes accelerated literature searches, analysis and prediction of physical and quantum chemical properties, transition states, chemical structures, chemical reactions, and also new catalysts and drug candidates.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

The Coimbra Chemistry Centre (CQC) was supported by the Fundação para a Ciência e a Tecnologia (FCT), Portugal, through Project UID/QUI/00313/2019.

ACKNOWLEDGMENTS

We thank all the authors who participated in this collection of articles, as well as all the reviewers for their relevant contributions to the Research Topic.

REFERENCES

Alexandrova, A. N., and Boldyrev, I. (2005). Search for the $Li_n^{0/+1/-1}$ ($n = 5$–7) lowest-energy structures using the *ab Initio* gradient embedded genetic algorithm (GEGA). Elucidation of the chemical bonding in the lithium clusters. *J. Chem. Theory Comput.* 1, 566–580. doi: 10.1021/ct050093g

Dieterich, J. M., and Hartke, B. (2017). *OGOLEM: Framework for GA-Based Global Optimization*. Available online at: https://www.ogolem.org/ (accessed March, 2020).

Hartke, B. (1993). Global geometry optimization of clusters using genetic algorithms. *J. Phys. Chem.* 97, 9973–9976. doi: 10.1021/j100141a013

Johnston, R. L. (2003). Evolving better nanoparticles: genetic algorithms for optimising cluster geometries. *Dalton Trans.* 4193. doi: 10.1039/B305686D

Llanio-Trujillo, J. L., Marques, J. M. C., and Pereira, F. B. (2011). An evolutionary algorithm for the global optimization of molecular clusters: application to water, benzene, and benzene cation. *J. Phys. Chem. A* 115, 2130–2138. doi: 10.1021/jp1117695

Marques, J. M. C., and Pereira, F. B. (2011). *EA_MOL: Evolutionary Algorithm for the Global Minimum Search of Molecular Clusters*. Available online at: https://apps.uc.pt/mypage/faculty/qtmarque/en/software/ (accessed February, 2020).

Martínez-Núñez, E. (2015a). An automated method to find transition states using chemical dynamics simulations. *J. Comput. Chem.* 36, 222–234. doi: 10.1002/jcc.23790

Martínez-Núñez, E. (2015b). An automated transition state search using classical trajectories initialized at multiple minima. *Phys. Chem. Chem. Phys.* 17:14912. doi: 10.1039/C5CP02175H

Martínez-Núñez, E., Barnes, G. L., Glowacki, D. R., Kopec, S. Pelaez-Ruiz, D. Rodriguez, A. et al. (2020). *AutoMeKin: Automated Mechanisms and Kinetics*. Available online at: https://rxnkin.usc.es/index.php/AutoMeKin (accessed January, 2020).

Ohno, K., and Maeda, S. (2006). Global reaction route mapping on potential energy surfaces of formaldehyde, formic acid, and their metal substituted analogues. *J. Phys. Chem. A* 110, 8933–8941. doi: 10.1021/jp061149l

Ohno, K., and Maeda, S. (2019). *Distribution of GRRM*. Available online at: https://iqce.jp/GRRM/index_e.shtml (accessed February, 2020).

Rodríguez-Fernández, R., Pereira, F. B., Marques, J. M. C., Martínez-Núñez, E., and Vázquez,. S. A. (2017). GAFit: a general-purpose, user-friendly program for fitting potential energy surfaces based on genetic algorithms. *Comput. Phys. Commun.* 217:89. doi: 10.1016/j.cpc.2017.02.008

Rodríguez-Fernández, R., Pereira, F. P., Marques, J. M. C., Vázquez-Rodríguez, S., and Martinez-Nunez, E. (2020). *GAFit, version 1.6*. Available online at: https://rxnkin.usc.es/index.php/GAFit (accessed January, 2020).

Shayeghi, A., Götz, D., Davis, J. B. A., Schäfer, R., and Johnston, R. L. (2015). Pool-BCGA: a parallelised generation-free genetic algorithm for the *ab initio* global optimisation of nanoalloy clusters. *Phys. Chem. Chem. Phys.* 17, 2104–2112. doi: 10.1039/C4CP04323E

Wales, D. J. (2010). *GMIN: A Program for Finding Global Minima and Calculating Thermodynamic Properties From Basin-Sampling*. Available online at: http://www-wales.ch.cam.ac.uk/GMIN/ (accessed January, 2020).

Wales, D. J., and Scheraga, H. A. (1999). Global optimization of clusters, crystals, and biomolecules. *Science* 285, 1368–1372. doi: 10.1126/science.285.5432.1368

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 12 July 2019
doi: 10.3389/fchem.2019.00485

Modified Particle Swarm Optimization Algorithms for the Generation of Stable Structures of Carbon Clusters, C_n ($n = 3$–6, 10)

Gourhari Jana[1], *Arka Mitra*[2], *Sudip Pan*[3], *Shamik Sural*[4]* *and Pratim K. Chattaraj*[1,5]*

[1] Department of Chemistry and Centre for Theoretical Studies, Indian Institute of Technology, Kharagpur, India, [2] Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology, Kharagpur, India, [3] Fachbereich Chemie, Philipps-Universität Marburg, Marburg, Germany, [4] Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, India, [5] Department of Chemistry, Indian Institute of Technology Bombay, Mumbai, India

Particle Swarm Optimization (PSO), a population based technique for stochastic search in a multidimensional space, has so far been employed successfully for solving a variety of optimization problems including many multifaceted problems, where other popular methods like steepest descent, gradient descent, conjugate gradient, Newton method, etc. do not give satisfactory results. Herein, we propose a modified PSO algorithm for unbiased global minima search by integrating with density functional theory which turns out to be superior to the other evolutionary methods such as simulated annealing, basin hopping and genetic algorithm. The present PSO code combines evolutionary algorithm with a variational optimization technique through interfacing of PSO with the Gaussian software, where the latter is used for single point energy calculation in each iteration step of PSO. Pure carbon and carbon containing systems have been of great interest for several decades due to their important role in the evolution of life as well as wide applications in various research fields. Our study shows how arbitrary and randomly generated small C_n clusters ($n = 3$–6, 10) can be transformed into the corresponding global minimum structure. The detailed results signify that the proposed technique is quite promising in finding the best global solution for small population size clusters.

Keywords: global minimum energy structures, density functional theory, carbon clusters, particle swarm optimization, multi-threaded code, Metaheuristic Algorithm, Gaussian

OPEN ACCESS

Edited by:
Jorge M. C. Marques,
University of Coimbra, Portugal

Reviewed by:
Wensheng Cai,
Nankai University, China
Nuno Lourenço,
University of Coimbra, Portugal
Alvaro Posada-Amarillas,
University of Sonora, Mexico

****Correspondence:***
Shamik Sural
shamik@cse.iitkgp.ac.in
Pratim K. Chattaraj
pkc@chem.iitkgp.ac.in
orcid.org/0000-0002-5650-7666

Specialty section:
This article was submitted to Theoretical and Computational Chemistry, a section of the journal Frontiers in Chemistry

Received: *09 April 2019*
Accepted: *24 June 2019*
Published: *12 July 2019*

Citation:
Jana G, Mitra A, Pan S, Sural S and Chattaraj PK (2019) Modified Particle Swarm Optimization Algorithms for the Generation of Stable Structures of Carbon Clusters, C_n ($n = 3$–6, 10). Front. Chem. 7:485. doi: 10.3389/fchem.2019.00485

INTRODUCTION

Over the past decades, studies on nature-inspired swarm intelligence based meta-heuristic algorithms have become a topic of paramount interest in the allied research fields. To date, various optimization problems have been addressed using these algorithms and these have turned out to be an important tool in analyzing physical systems, in solving the complex problems and in searching for the best solution from a set of all possible feasible solutions. Particularly, global optimization (GO) has become very challenging in the development of computational fields. Search for the globally optimal solution is more crucial than that for a local optima as the former corresponds to the correct and desirable solution. Fundamentally, GO methods can be divided into two broad classes, namely (i) deterministic algorithms and (ii) stochastic algorithms. Although deterministic methods are capable of providing a guaranteed global optimum solution, the necessary properties of objective function and some constraints are required as well. On the other hand, stochastic methods can provide successful results in finding the global best solution without consideration of any assumption of differentiability and continuity of objective function. Until now, several stochastic methods such as genetic algorithms (GA) (Holland, 1992; Grüninger and Wallace, 1996; Ursem, 2000; Deb et al., 2002; Poli and Langdon, 2002; Dilettoso and Salerno, 2006; Krug et al., 2010), simulated annealing (SA) (Woodley et al., 1999; Abraham and Probert, 2006; Glass et al., 2006; Oganov and Glass, 2006; Trimarchi and Zunger, 2007), differential evolution (DE) (Storn, 1996; Storn and Price, 1997; Price et al., 2006; Rocca et al., 2011), harmony search (HS) (Geem, 2000, 2001, 2006; Geem et al., 2001, 2005; Diao and Shen, 2012; Gholizadeh and Barzegar, 2013; Hadwan et al., 2013; Manjarres et al., 2013; Nekooei et al., 2013; Wang and Li, 2013; Hoang et al., 2014; Fattahi et al., 2015; Weyland, 2015; Assad and Deep, 2016), ant colony optimization (ACO) (Colorni et al., 1992; Dorigo, 1992; Dorigo and Di Caro, 1999; Zlochin et al., 2004; Dorigo and Birattari, 2010; Korošec et al., 2012), cuckoo search (CS) (Payne and Sorensen, 2005; Yang and Deb, 2009; Inderscience, 2010), bat algorithm (BA) (Altringham et al., 1996; Richardson, 2008; Yang, 2010a,b), artificial bee colony optimization (ABC) (Karaboga and Basturk, 2007, 2008; Omkar et al., 2011; Fister and Žumer, 2012; Li G. et al., 2012), honey bee mating optimization (HBMO); (Pham et al., 2005; Haddad et al., 2006; Afshar et al., 2007; Jahanshahi and Haddad, 2008; Marinakis and Marinaki, 2009; Pham and Castellani, 2009, 2014, 2015; Bitam et al., 2010; Gavrilas et al., 2010; Marinaki et al., 2010; Chakaravarthy and Kalyani, 2015; Nasrinpour et al., 2017; Rajasekhar et al., 2017), and multi-colony bacteria foraging optimization (MC-BFA) (Chen et al., 2010) have been developed and used in various research fields including global optimization purpose. Moreover, some advanced and more promising methods are continuously being proposed including random sampling method (Pickard and Needs, 2006, 2007, 2008), minima hopping (Kirkpatrick et al., 1983; Pannetier et al., 1990), basin hopping (Nayeem et al., 1991; Wales and Doye, 1997), meta-dynamics (Martonák et al., 2003, 2005; Guangneng et al., 2005), data mining (Mujica and Needs, 1997) and Particle Swarm Optimization (PSO) (Kennedy and Eberhart, 1995, 1999; Kennedy, 1997; Shi and Eberhart, 1998; Eberhart and Shi, 2001; Li, 2007; Özcan and Yilmaz, 2007; Poli, 2007, 2008; Barrera and Coello, 2009; Li M. et al., 2012; Qu et al., 2012; Bonyadi and Michalewicz, 2017), modified PSO (Zheng et al., 2007), adaptive particle swarm optimization (APSO) (Zhan et al., 2009), multi-dimensional PSO for dynamic environments (Zhi-Jie et al., 2009; Kiranyaz et al., 2011; Bhushan and Pillai, 2013), which indeed show different numerical performances.

Out of these numerous techniques, PSO is a very renowned iterative process which works intelligently by utilizing the concept of exploring and exploiting together in the multidimensional search space for finding optimal or near-optimal solutions. The learning strategies of this technique for the searching of structural information are very much suitable and reliable in an active area of GO research. This evolutionary computational method was first invented by Kennedy and Eberhart (1995) and Kennedy (1997) in the mid 1990s on graceful collaborative motion of biological populations rooted on the concept of "information sharing and collective intelligence." This adaptive metahurestic technique emphasizes on overcoming the energy barriers, particularly by the upgradation of positions and velocities following the individual or personal best which again follows the global best one. After several developments (Reeves, 1983; Reynolds, 1987; Heppner and Grenander, 1990; Millonas, 1993; Clerc, 1999; Eberhart and Shi, 2000; Banks et al., 2007; Bui et al., 2007; Khan and Sadeequllah, 2010), adaptation (Wang et al., 2011), modifications (like niching with PSO Brits et al., 2002; Engelbrecht and Van Loggerenberg, 2007; Sun et al., 2007; Nickabadi et al., 2008; Wang J. et al., 2009; Wang Y. et al., 2009) single solution PSO (Liu and Wang, 2006; AlRashidi and El-Hawary, 2007; Li and Li, 2007; Liu B. et al., 2007; Liu D. et al., 2007; Petalas et al., 2007; Schutze et al., 2007; Zhang et al., 2007; Zhang and Wang, 2008; Benameur et al., 2009) and multi-objective optimization (Cai et al., 2004, 2009; Call et al., 2007; Chandrasekaran et al., 2007; Abido, 2009; Alatas and Akin, 2009; Dehuri and Cho, 2009; De Carvalho et al., 2010; Goh et al., 2010; Briza and Naval, 2011; Chen et al., 2011), constraint optimization with PSO (Cao et al., 2004; AlRashidi and El-Hawary, 2006; Sun and Gao, 2008; Ma et al., 2009; Sivasubramani and Swarup, 2009), discrete PSO (Yin, 2004; Yeh, 2009; Yeh et al., 2009; Unler and Murat, 2010), dynamic environment of PSO (Shao et al., 2004, 2008; Zhang et al., 2006; Chen et al., 2007; Liu X. et al., 2007; Yang et al., 2007; Du and Li, 2008; Wang and Xing, 2008; Zhao et al., 2008; Cheng et al., 2009; Wang Y. et al., 2009; Bae et al., 2010) and parameterization (Eberhart and Shi, 2001; Shi, 2001; Trelea, 2003; Li-Ping et al., 2005; Talbi, 2009; Pedersen, 2010; Bansal et al., 2011) on the original PSO, more recently global optimization of small boron clusters (B_5 and B_6) using a more advanced PSO approach has been reported with great success (Mitikiri et al., 2018).

On the other hand, the investigation on pure carbon molecules existing in various structural forms (chains/cyclic rings) has been a matter of great interest in the research area of organic, inorganic and physical chemistry (Weltner and Van Zee, 1989) as the study and production of carbon-riched molecules in the laboratory are notoriously difficult due to their high reactivity and transient like behavior. They are also very important in astrophysics, particularly in connection with the chemistry of carbon stars (Bernath et al., 1989), comets (Douglas, 1951), and interstellar molecular clouds (Bettens and Herbst, 1997). Long carbon chains are also believed to act as carriers of diffuse interstellar bands (Fulara et al., 1993). Moreover, carbon clusters are also important constituents in hydrocarbon flames and other soot-forming systems (Kroto and McKay, 1988) and they play an important role in gas-phase carbon chemistry where they serve as intermediates for the production of fullerenes, carbon tubes, thin diamond and silicon carbide films (Koinuma et al., 1996; Van Orden and Saykally, 1998). Therefore, the study about the structures and stabilities of carbon clusters is very important to thoroughly understand the complex chemical environment of such systems and also to shed light into the remarkable bonding capability of carbon which is able to form single, double and triple bonds. They together make the study on the structural information of carbon clusters in the field of theoretical research a subject of immense interest and it started before the development of fullerene chemistry (Pitzer and

Clementi, 1959; Weltner and Van Zee, 1989; Martin et al., 1993; Hutter et al., 1994).

Due to the reduction in angle strain, carbon clusters larger than C_{10} are likely to exist as monocyclic rings, while smaller ones possess low-energy linear structures. Moreover, it was reported that for small clusters with even number of carbon atoms such as C_4, C_6, and C_8, the cyclic form is either the lowest energy isomer or almost isoenergetic to their linear counterparts (Raghavachari and Binkley, 1987; Watts et al., 1992; Hutter and Lüthi, 1994; Pless et al., 1994; Martin and Taylor, 1996). In this study, we have checked the efficiency of our newly developed multi-threaded PSO code, written in python, and augmented by Gaussian 09 program package (Frisch et al., 2013) to locate global minimum energy structures for C_n clusters ($n = 3$–6). Particularly, we want to test our code for the system where two minima are located at two deep well points on the PES as in the case of C_6 cluster. We kept the cluster size small in order to compare the performance of our code to other popular evolutionary simulation techniques such as SA, GA, and BH.

CURRENTLY PROPOSED AND IMPLEMENTED PSO TECHNIQUE

Initially, random structures are generated within certain range (−3, 3) in a multidimensional search space followed by upgradation of velocity and position vectors through swarm intelligence. After completion of every iteration, energy of each particle is calculated and a convergence criterion is verified with the help of the Gaussian 09 package interfaced with the present PSO algorithm. Individual best and global best positions are updated. If the energy values of successive 30 iterations remain same, the program automatically terminates. Finally a new set of initial structures are generated from the related output structures and the process is continued till the self-consistency is achieved.

In order to check the efficiency of our proposed PSO method over some most familiar GO methods like advanced BH, SA, and GA methods, the results for C_5 cluster have been analyzed, as a reference system.

A COMPARATIVE ACCOUNT OF THE CURRENT PSO METHOD WITH OTHER EXISTING APPROACHES

We have made the computer experiment to compare our proposed PSO with the other popular evolutionary simulation techniques such as SA, GA and advanced BH.

Comparison of Performances of PSO and GA

(a) The most important distinction between our proposed DFT-PSO with GA is the sharing of information. In GA, chromosomes share information with each other, whereas in PSO the best particle informs the others and the information of variables is stored in small memory. Again, PSO search for the global best solution is unidirectional, while GA follows the parallel searching process.

(b) In contrast to GA, PSO does not use any genetic kind of operator, i.e., crossover and mutation, and the internal velocity leads the particle to the next better place.

(c) PSO implementation is more simple and easier than GA as it deals with few parameters (like position and velocity only).

(d) GA provides satisfactory results in case of combinatorial problems, PSO being less suitable there.

(e) PSO takes much less time to execute and the convergence rate is also faster than that of GA.

A previous study by Hassan et al. (2005) has been further recommended for more clarity and reliability of the efficiency of PSO over GA.

Comparison of Performances of PSO and SA

In SA technique, a small perturbation is given to cluster entity at each successive step, and energy estimation is carried out consecutively. Acceptance of perturbation depends on the obtained energy value. If the obtained energy is better than the previous one, the perturbation as well as the move with low cost is accepted. Otherwise, the process excludes it and the Boltzmann probability distribution is applied at a given temperature. The particle (individual cluster) in SA takes much time to generate different lower energy structures. The temperature decreases during the whole course of the process very slowly and at the end of the run it attains the least value. In contrast, such kind of perturbation or temperature variable is not present in PSO. Both exploitation and exploration techniques drive the particle in PSO, while only exploitation is used in SA. So, there are more chances to trap the particles in local minima in case of SA being a single-based technique than PSO. On the other hand, PSO, being the population based technique, is able to swarm wherever (different places of mountain or lower point of valleys) be the particle in the search space.

Comparison With Basin Hopping

Wales and Doye jointly described basin hopping algorithm (Berg and Neuhaus, 1991; Wales and Doye, 1997; Doye et al., 1998) which has become a popular stochastic search process to find out the desired global best solution of an object function. This method is basically a Monte Carlo technique, which works in a perturbative and iterative manner. At first, a random coordinate of a particle is considered. Then, random perturbation is applied to the configuration considering the fact that the particle remains in a local basin which is then followed by the minimization of energy functional to get a better solution. Energy estimation is again carried out and the process is repeated until the best configuration or the lowest energy structure is achieved. The most important thing is that the applied perturbation should be large enough to get out of a local basin.

ALGORITHM AND COMPUTATIONAL DETAILS

At the beginning, a set of random coordinates of C_n clusters (particles) with random positions and velocities are considered.

The newer sets of coordinates are updated through PSO run to find out global best position or configuration. The local best configuration (p_{best}) or that having the lowest energy value obtained locally so far is stored in a small memory variable which is then followed by the searching of global best (g_{best}) configuration (among the set of p_{best}) through an exploration technique. Ultimately, the best optimal solution is achieved.

The new velocities ($v_{\mathrm{i}}^{\mathrm{t+1}}$) and positions ($x_{\mathrm{i}}^{\mathrm{t+1}}$) of particles in ith generation obey the following equations where $x_{\mathrm{i}}^{\mathrm{t}}$ and $v_{\mathrm{i}}^{\mathrm{t}}$ are the current position and velocity.

$$v_{\mathrm{i}}^{\mathrm{t+1}} = w * v_{\mathrm{i}}^{\mathrm{t}} + d_1 * \varepsilon_1 * (p_{best} - x_{\mathrm{i}}^{\mathrm{t}}) + d_2 * \varepsilon_2 * (g_{best} - x_{\mathrm{i}}^{\mathrm{t}}) \quad (1)$$

$$x_{\mathrm{i}}^{\mathrm{t+1}} = x_{\mathrm{i}}^{\mathrm{t}} + v_{\mathrm{i}}^{\mathrm{t+1}} \quad (2)$$

ε_1 and ε_2 are chosen randomly in between (0,1). The tendency of a particle to remain in its current position is called inertia coefficient denoted by w. d_1 and d_2 (which can be modified as per requirement) which are referred to as individual coefficient of acceleration and global coefficient of acceleration, respectively. These two coefficients guide the particles to meet convergence so that all the candidate solutions in the problem space efficiently achieve the global minimum (see **Table 1**).

TABLE 1 | PSO Parameters.

Parameters	Value
Population (N_{pop})	10
Inertia Coefficient (w)	0.4–0.8
Individual coefficient of acceleration (d_1)	2
Global coefficient of acceleration (d_2)	2
Random Coefficients (ε_1; ε_2)	[0,1]

After the completion of each PSO run, optimization of global best structural units of C_n clusters (n = 3–6) are performed at the B3LYP (Lee et al., 1988; Becke, 1993)/6-311+G* level in the Gaussian 09 program.

Each randomly generated cluster unit is considered as a particle. In **Figure 1** (x_0, x_1, x_2,… x_{3n-1}), particle comprises n atoms. Here, the coordinates of ith atom are (x_{3i}, x_{3i+1}, x_{3i+2}).

PARALLEL IMPLEMENTATION

One of the major advantages of using PSO as proposed in this paper over some of the classical optimization techniques is its parallelizability. The same implementation of the algorithm can be executed on machines having single core (serial implementation) or ones with multiple cores (laboratory grade clusters) or high performance computing (HPC) systems. Changing a couple of header parameters in the program is sufficient to make it portable across a wide range of platforms. We have tested both a serial as well as a parallel implementation of our programs. Results on parallel implementation are reported. It may be noted that our PSO algorithm implemented in Python invokes the Gaussian software as a system call. Each such parallel call, one for each particle of the PSO algorithm, causes a new instance of Gaussian to be executed. The number of cores on which each Gaussian instance runs is dependent on the available number of processor cores. However, at the

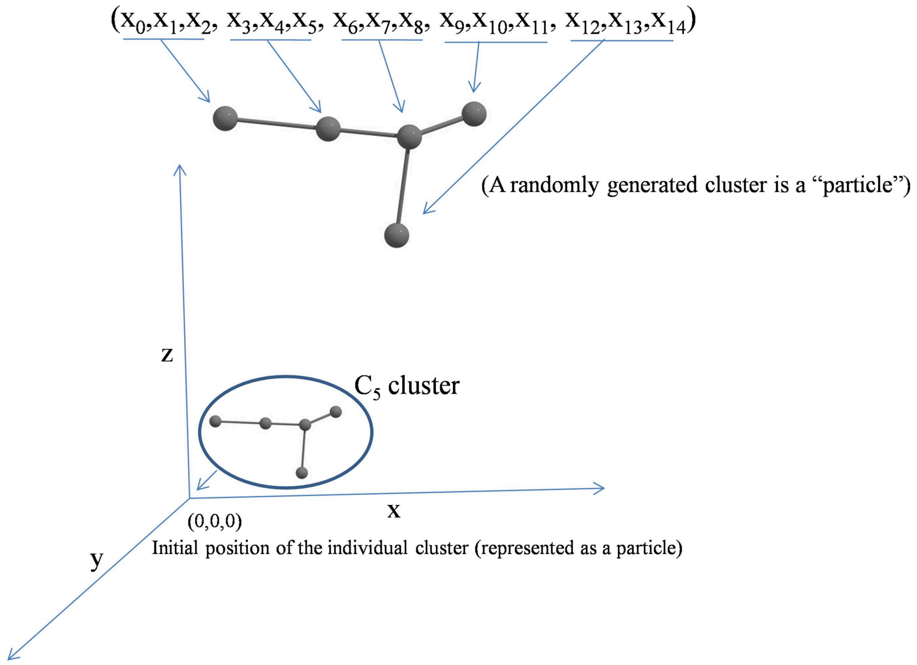
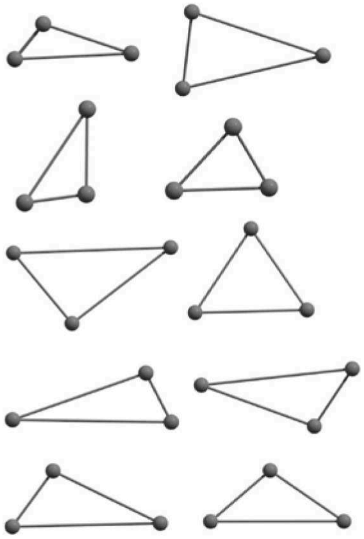
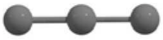
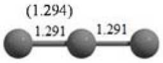
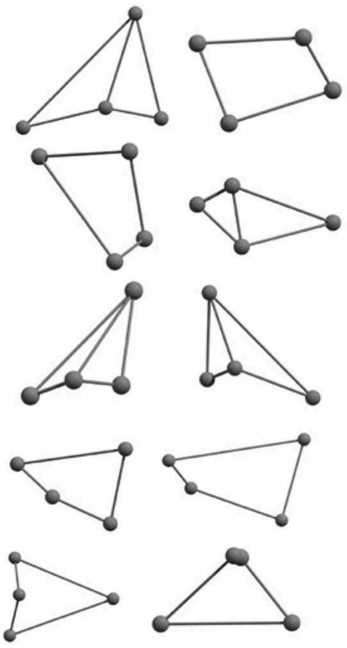
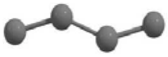
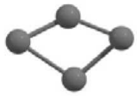
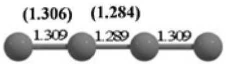
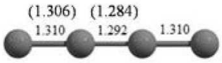
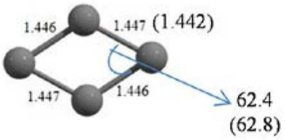

(x_0,x_1,x_2, x_3,x_4,x_5, x_6,x_7,x_8, x_9,x_{10},x_{11}, x_{12},x_{13},x_{14})
(A randomly generated cluster is a “particle”)
z
C_5 cluster
x
(0,0,0)
Initial position of the individual cluster (represented as a particle)
y

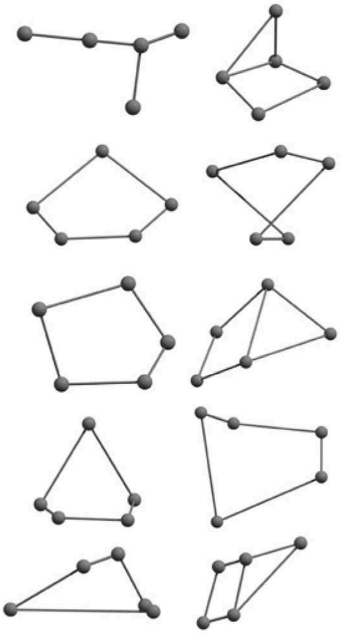
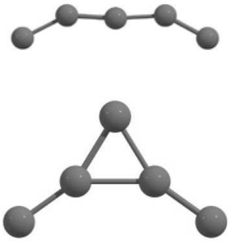
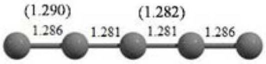
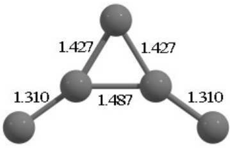
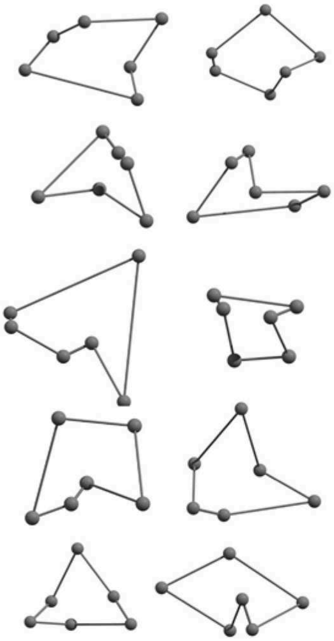
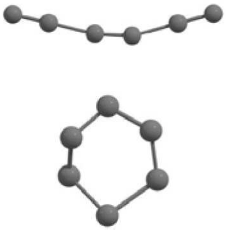
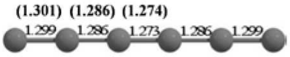
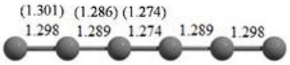
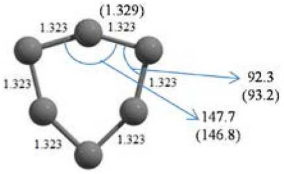
FIGURE 1 | A schematic representation of a cluster in multidimensional search space.

TABLE 2 | The randomly chosen 10 different molecular frameworks of C_n (n = 3–6, 10) with singlet and triplet spin multiplicity converge to the global minimum energy structures (Bond lengths are given in Å unit and the relative energies, ΔE w.r.t the global minimum energy structures in brackets are given in kcal/mol).

Clusters	Initial structure	Final structure using PSO	Final optimized energy (bond lengths)
C_3 cluster			 $D_{\infty h}$, S (E = −114.0769 a.u.)
C_4 cluster		 	 $D_{\infty h}$, T E = −152.1320 a.u. [0.0]  $D_{\infty h}$, S E = −152.1036 a.u. [17.8]  1.446 1.447 (1.442) 1.447 1.446 62.4 (62.8) D_{2h}, S E = −152.1062 a.u. [16.2]

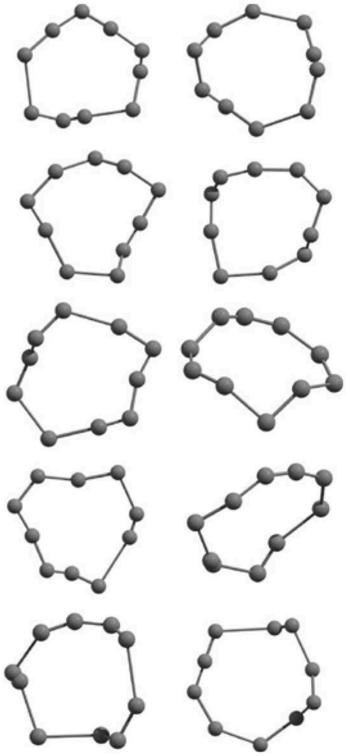
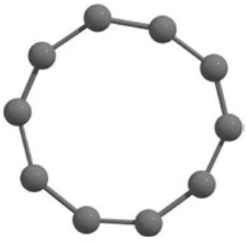
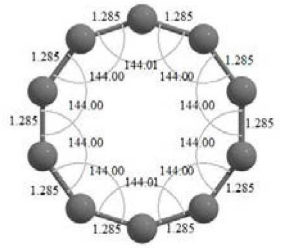
(Continued)

TABLE 2 | Continued

Clusters	Initial structure	Final structure using PSO	Final optimized energy (bond lengths)
C_5 cluster			 $D_{\infty h}$, S E = −190.2546 a.u. [0.0]  C_{2v}, S E = −190.1350 a.u. [75.1]
C_6 cluster			 (1.301) (1.286) (1.274) 1.299 1.286 1.273 1.286 1.299 $D_{\infty h}$, T E = −228.3181 a.u. [0.0]  (1.301) (1.286) (1.274) 1.298 1.289 1.274 1.289 1.298 $D_{\infty h}$, S E = −228.2969 a.u. [13.3]  (1.329) 1.323 1.323 1.323 1.323 1.323 1.323 92.3 (93.2) 147.7 (146.8) D_{3h}, S E = −228.3071 a.u. [6.9]

(Continued)

TABLE 2 | Continued

Clusters	Initial structure	Final structure using PSO	Final optimized energy (bond lengths)
C_{10} cluster			 D_{10h}, S (E = −380.7543 a.u.)

All coordinates are provided in ***Supporting Information****. (Experimental bond lengths and angles are provided within the parenthesis in the final optimized structure).*

end of every iteration, PSO has to recompute the best and global best positions of individual particle before updating the velocity values from which the new positions of the particles are determined. These are done by reading the output log files generated by Gaussian for each particle. This implies that the results of all the parallel invocations of Gaussian need to be completed before the iteration-end processing can be done. We have implemented appropriate synchronization mechanisms to enable such parallel implementation and hence, the code base is portable across multiple platforms.

COMPUTATIONAL SETUP

All our computations were carried out on a single server having two Intel 2.70 GHz Xeon E5-2697 v2 processors and 256 GB of RAM. Each processor has 12 cores. Leaving aside a few cores for operating system and other housekeeping processes, we made use of 30 threads for executing our PSO algorithm. A PSO population size of 15 particles implies that 2 threads could be used for each instance of Gaussian. Also, 8 GB of RAM was dedicated to each such instance. As mentioned before, the number of PSO particles, RAM assignment and the number of threads for each Gaussian call are set as input hyper parameters. The completely parameterized implementation of PSO has been done in Python invoking Gaussian for energy calculation in a multi-threaded environment. This is one of the unique features of our work, which has not yet been reported in the literature for stable structure prediction of C_n, to the best of our knowledge.

RESULTS AND DISCUSSION

In our study, each C_n cluster unit (each individual unit) is considered to be a swarm particle in a multidimensional potential energy surface (PES) where the stationary points (maxima, minima, and higher order saddle points) are connected. The randomly generated individual particle is governed by a position vector and a velocity vector. Again, each position vector representing a candidate solution in the hyperspace starts searching for the optima of a given function of several variables by updating generations in iterative process without much of any assumption leading to a minimum energy structure. After iteration the particle driven by a velocity vector changes its search direction. The position and velocity vectors together store the information regarding its own best position or the local best position (called p_{best}) seen so far and a global best position (called g_{best}) which is obtained by communicating with its nearest neighbors. Further, the advancement of particles toward the global best position is attained via particle swarm optimizer ideology and they gravitate toward the global best solution with the help of the best variable memory values. Our proposed PSO implementation explores rapidly without being entrapped in local optima and executes extensively, followed by immediate convergence to the desired objective value, the global optima.

The results of global optimization of C_n clusters ($n = 3$–6, 10) considering a maximum of 1,000 runs starting from the random choices of input configuration are shown in **Table 2**. The global stable structure (best solution) can be obtained by fulfilling the termination criteria along the convex function of the information matrix when one of the particles reaches the target. Initially, 10 different random configurations have been chosen by setting random initial positions and velocities of all particles followed by the Gaussian interfaced PSO driven operation to get the global optimum structure (see **Table 2**).

It is a very fascinating aspect that Gaussian optimization technique works in such a way that the guess structure can be stuck at local minima which may or may not be the global minimum. But, it is obvious that our proposed modified PSO implementation converges to the most stable structure where all the particles exist in a given range in the multidimensional hyperspace. However, sometimes atoms of the randomly generated particles (each individual cluster unit) are not in the limit of bonding perception and they might overlap on each other. In order to understand whether the atoms remain in the same molecular framework or not, we have connected the randomly deployed particles with solid lines in the following figures and they do not necessarily imply true bonds (see **Table 2**). In case of C_3 cluster, the structure obtained after the end of the PSO run (linear, $D_{\infty h}$ point group) exactly matches with the structure obtained after the final G09 optimization in terms of bond length and energy. C_5 cluster also shows linear geometry with $D_{\infty h}$ point group and singlet electronic state after final optimization step. A significantly higher energy cyclic isomer is also found in this case. On the other hand, C_4 and C_6 clusters (containing even number of C atoms) give both linear ($D_{\infty h}$) and ring structures (D_{2h} for C_4 and D_{3h} for C_6). Corresponding energies and bond lengths are provided in **Table 2**. The computed geometrical parameters and minimum energy structures match excellently with the previously reported experimental results (Raghavachari and Binkley, 1987; Watts et al., 1992; Hutter and Lüthi, 1994; Pless et al., 1994; Martin and Taylor, 1996; Van Orden and Saykally, 1998). For both C_4 and C_6 clusters, the lowest energy isomer has linear form in triplet state, whereas the linear singlet state is 17.8 (C_4) and 13.3 (C_6) kcal/mol higher in energy than the corresponding triplet forms. In addition to the small cluster systems, we have also checked the efficiency and the robustness of our implemented PSO code to find the global

TABLE 3 | Comparison of PSO results with other more popular evolutionary GO techniques as applied to the C_5 cluster starting from the corresponding local minima structures.

Comparison in terms of	Advanced basin hopping (BH)	Simulated annealing (SA)	Modified PSO
Execution time to locate the global minimum (GM)	305,140 s	12,959 s	8,898 s
Energy of the global minimum (Energy after completion of iterations)	−190.2546 a.u. (−190.2460 a.u.)	−190.2546 a.u. (−189.5141 a.u.)	−190.2546 a.u. (−190.2436 a.u.)
Number of iterations needed to get a structure close to GM	1,703 (converged)	92 (not converged)	331 (converged)

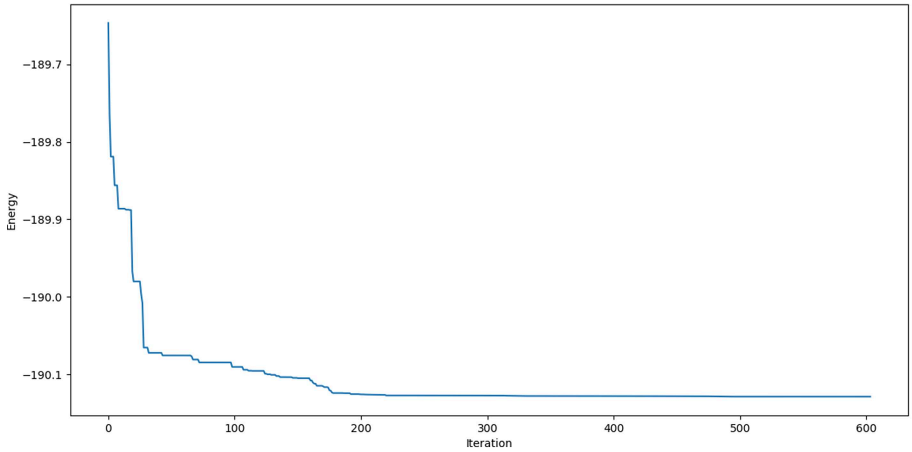


FIGURE 2 | Single point energy evolution landscape of C_5 cluster during each generation of convergence at the B3LYP/6-311+G* level.

minimum for a relatively larger sized cluster, C_{10}. The results show that the present code can successfully locate the desired D_{10h} symmetric ring structure which is the most stable isomer in this case.

In the present context, we have also carried out DFT-SA and DFT-BH methods considering same object energy function as in our proposed PSO method to compare the obtained results (see **Table 3**). The tabulated values clearly reflect that the present PSO method is superior to other methods based on the time to locate the GM, energy values after completion of all runs of the studied methods and the number of iteration steps needed to get the final structure.

A representative plot of C_5 cluster (as reference) is given below to ensure the fulfillment of convergence criteria up to 600 iteration steps (see **Figure 2**).

CONCLUSION

This systematic study for the searching of the most stable carbon based small clusters describes the effectiveness of the application of our proposed PSO technique. Currently employed less expensive and relatively less complicated computational method generates a vast potential search space depending only on the position and velocity variables. Our proposed method opens a new vista to find out global minimum energy structures effectively and accurately within a given multidimensional configuration search space. PSO implementation without much of any assumption like constraints of symmetry and externally imposed factors like temperature, pressure, etc. performs suitably and converges to a single configuration that presumably is a global minimum energy structure or may exactly fit it after Gaussian optimization. PSO can be used as a fast post-processing technique to get a global minimum or close to global minimum structure. In fact, in this study we have introduced a new easy implementation and computationally less expensive approach for the reduction of iteration steps to obtain global best configurations of small carbon clusters with exact energy values.

DATA AVAILABILITY

All datasets generated for this study are included in the manuscript and/or the **Supplementary Files**.

AUTHOR CONTRIBUTIONS

GJ: interfacing with the Gaussian software, writing first version of the full manuscript, generation of the TOC, and analysis of the results. AM: implementation of the PSO algorithm in python including coding and parallelization. SP: revision of the draft manuscript and interpretation of data. SS: revision of the draft manuscript and checking of implemented code. PC: formulation of the problem, critical revision of the manuscript, and data interpretation.

FUNDING

This work was supported by the Faculty Development Fund (FDF) (SS) and Department of Science and Technology, Ministry of Science and Technology [Award Number: JBF (PC) Grant number-SR/S2/JCB-09/2009].

ACKNOWLEDGMENTS

PC thanks Professors Jorge M. C. Marques, Emilio Martinez-Nunez, and William Hase for kindly inviting him to contribute to this Special Issue of Frontiers in Chemistry, on Application of Optimization Algorithms in Chemistry. He would also like to thank DST, New Delhi for the J. C. Bose National Fellowship. SS and AM thank CSE for the computational facilities. GJ thanks IIT, Kharagpur for his fellowship.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2019.00485/full#supplementary-material

REFERENCES

Abido, M. (2009). Multiobjective particle swarm optimization for environmental/economic dispatch problem. *Electr. Pow. Syst. Res.* 79, 1105–1113. doi: 10.1016/j.epsr.2009.02.005

Abraham, N. L., and Probert, M. I. (2006). A periodic genetic algorithm with real-space representation for crystal structure and polymorph prediction. *Phys. Rev. B* 73:224104. doi: 10.1103/PhysRevB.73.224104

Afshar, A., Haddad, O. B., Mariño, M. A., and Adams, B. J. (2007). Honey-bee mating optimization (HBMO) algorithm for optimal reservoir operation. *J. Franklin Inst.* 344, 452–462. doi: 10.1016/j.jfranklin.2006.06.001

Alatas, B., and Akin, E. (2009). Multi-objective rule mining using a chaotic particle swarm optimization algorithm. *Knowl Based Syst.* 22, 455–460. doi: 10.1016/j.knosys.2009.06.004

AlRashidi, M., and El-Hawary, M. (2007). Hybrid particle swarm optimization approach for solving the discrete OPF problem considering the valve loading effects. *IEEE Trans. Power Syst.* 22, 2030–2038. doi: 10.1109/TPWRS.2007.907375

AlRashidi, M. R., and El-Hawary, M. E. (2006). "Emission-economic dispatch using a novel constraint handling particle swarm optimization strategy," in *Electrical and Computer Engineering, CCECE'06. Canadian Conference on IEEE* (Ottawa, ON: IEEE), 664–669.

Altringham, J., McOwat, T., and Hammond, L. (1996). *Bats: Biology and Behaviour*. New York, NY: Oxford University Press.

Assad, A., and Deep, K. (2016). "Applications of harmony search algorithm in data mining: a survey," in *Proceedings of Fifth International Conference on Soft Computing for Problem Solving* (Singapore: Springer), 863–874.

Bae, C., Yeh, W.-C., Chung, Y. Y., and Liu, S.-L. (2010). Feature selection with intelligent dynamic swarm and rough set. *Expert Syst. Appl.* 37, 7026–7032. doi: 10.1016/j.eswa.2010.03.016

Banks, A., Vincent, J., and Anyakoha, C. (2007). A review of particle swarm optimization. Part I: background and development. *Nat. Comput.* 6, 467–484. doi: 10.1007/s11047-007-9049-5

Bansal, J. C., Singh, P., Saraswat, M., Verma, A., Jadon, S. S., and Abraham, A. (2011). "Inertia weight strategies in particle swarm optimization," in *Nature and Biologically Inspired Computing (NaBIC), Third World Congress on IEEE* (Salamanca: IEEE), 633–640.

Barrera, J., and Coello, C. C. A. (2009). “A particle swarm optimization method for multimodal optimization based on electrostatic interaction,” in *8th Mexican International Conference on Artificial Intelligence, MICAI 2009: Advances in Artificial Intelligence*, MICAI 2009, Lecture Notes in Computer Science, Vol. 5845, eds A. H. Aguirre, R. M. Borja, and C. A. R. Garci? (Berlin; Heidelberg: Springer), 622–632. doi: 10.1007/978-3-642-05258-3_55

Becke, A. D. (1993). Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* 98, 5648–5652. doi: 10.1063/1.464913

Benameur, L., Alami, J., and El Imrani, A. (2009). “A new hybrid particle swarm optimization algorithm for handling multiobjective problem using fuzzy clustering technique,” in *2009 International Conference on Computational Intelligence, Modelling and Simulation* (Brno: IEEE), 48–53. doi: 10.1109/CSSim.2009.42

Berg, B. A., and Neuhaus, T. (1991). Multicanonical algorithms for first order phase transitions. *Phys. Lett. B* 267, 249–253. doi: 10.1016/0370-2693(91)91256-U

Bernath, P. F., Hinkle, K. H., and Keady, J. J. (1989). Detection of C5 in the circumstellar shell of IRC+ 10216. *Science* 244, 562–564. doi: 10.1126/science.244.4904.562

Bettens, R., and Herbst, E. (1997). The formation of large hydrocarbons and carbon clusters in dense interstellar clouds. *Astrophys. J.* 478:585. doi: 10.1086/303834

Bhushan, B., and Pillai, S. S. (2013). “Particle swarm optimization and firefly algorithm: performance analysis,” in *2013 3rd IEEE International Advance Computing Conference* (Ghaziabad: IEEE).

Bitam, S., Batouche, M., and Talbi, E.-G. (2010). “A survey on bee colony algorithms,” in *IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW)* (Atlanta, GA: IEEE), 1–8.

Bonyadi, M. R., and Michalewicz, Z. (2017). *Particle Swarm Optimization for Single Objective Continuous Space Problems: A Review*. MIT Press, 25, 1–54. doi: 10.1162/EVCO_r_00180

Brits, R., Engelbrecht, A. P., and Van den Bergh, F. (2002). “A niching particle swarm optimizer,” in *Proceedings of the 4th Asia-Pacific Conference on Simulated Evolution and Learning* (Singapore: Orchid Country Club), 692–696.

Briza, A. C., and Naval, P. C. Jr. (2011). Stock trading system based on the multi-objective particle swarm optimization of technical indicators on end-of-day market data. *Appl. Soft Comput.* 11, 1191–1201. doi: 10.1016/j.asoc.2010.02.017

Bui, L. T., Soliman, O., and Abbass, H. (2007). “A modified strategy for the constriction factor in particle swarm optimization,” in *Progress in Artificial Life. ACAL 2007. Lecture Notes in Computer Science*, Vol. 4828, eds M. Randall, H. A. Abbass, and J. Wiles (Berlin; Heidelberg: Springer), 333–344. doi: 10.1007/978-3-540-76931-6_29

Cai, J., Ma, X., Li, Q., Li, L., and Peng, H. (2009). A multi-objective chaotic particle swarm optimization for environmental/economic dispatch. *Energy Convers. Manag.* 50, 1318–1325. doi: 10.1016/j.enconman.2009.01.013

Cai, W., Shao, N., Shao, X., and Pan, Z. (2004). Structural analysis of carbon clusters by using a global optimization algorithm with Brenner potential. *J. Mol. Struct. Theochem* 678, 113–122. doi: 10.1016/j.theochem.2004.03.017

Call, S. T., Zubarev, D. Y., and Boldyrev, A. I. (2007). Global minimum structure searches via particle swarm optimization. *J. Comput. Chem.* 28, 1177–1186. doi: 10.1002/jcc.20621

Cao, C.-H., Li, W.-H., Zhang, Y.-J., and Yi, R.-Q. (2004). “The geometric constraint solving based on memory particle swarm algorithm,” in *Machine Learning and Cybernetics, 2004. Proceedings of International Conference on: IEEE* (Shanghai: IEEE), 2134–2139.

Chakaravarthy, T., and Kalyani, K. (2015). A brief survey of honey bee mating optimization algorithm to efficient data clustering. *Indian J. Sci. Technol.* 8:24. doi: 10.17485/ijst/2015/v8i24/59219

Chandrasekaran, S., Ponnambalam, S., Suresh, R., and Vijayakumar, N. (2007). “Multi-objective particle swarm optimization algorithm for scheduling in flowshops to minimize makespan, total flowtime and completion time variance,” in *Evolutionary Computation, 2007. CEC 2007* (Singapore: IEEE Congress), 4012–4018.

Chen, H., Zhu, Y., and Hu, K. (2010). Multi-colony bacteria foraging optimization with cell-to-cell communication for RFID network planning. *Appl. Soft Comput.* 10, 539–547. doi: 10.1016/j.asoc.2009.08.023

Chen, H., Zhu, Y., Hu, K., and Ku, T. (2011). RFID network planning using a multi-swarm optimizer. *J. Netw. Comput. Appl.* 34, 888–901. doi: 10.1016/j.jnca.2010.04.004

Chen, Y.-P., Peng, W.-C., and Jian, M.-C. (2007). Particle swarm optimization with recombination and dynamic linkage discovery. *IEEE Trans. Syst. Man Cybern. B* 37, 1460–1470. doi: 10.1109/TSMCB.2007.904019

Cheng, C.-T., Liao, S.-L., Tang, Z.-T., and Zhao, M.-Y. (2009). Comparison of particle swarm optimization and dynamic programming for large scale hydro unit load dispatch. *Energy Convers. Manag.* 50, 3007–3014. doi: 10.1016/j.enconman.2009.07.020

Clerc, M. (1999). “The swarm and the queen: towards a deterministic and adaptive particle swarm optimization,” in *Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on: IEEE* (Washington, DC: IEEE), 1951–1957.

Colorni, A., Dorigo, M., and Maniezzo, V. (1992). “Distributed optimization by ant colonies,” in *Proceedings of the First European Conference on Artificial Life* (Cambridge, MA), 134–142.

De Carvalho, A. B., Pozo, A., and Vergilio, S. R. (2010). A symbolic fault-prediction model based on multiobjective particle swarm optimization. *J. Syst. Softw.* 83, 868–882. doi: 10.1016/j.jss.2009.12.023

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* 6, 182–197. doi: 10.1109/4235.996017

Dehuri, S., and Cho, S.-B. (2009). Multi-criterion Pareto based particle swarm optimized polynomial neural network for classification: a review and state-of-the-art. *Comput. Sci. Rev.* 3, 19–40. doi: 10.1016/j.cosrev.2008.11.002

Diao, R., and Shen, Q. (2012). Feature selection with harmony search. *IEEE Trans. Syst. Man Cybern. B* 42, 1509–1523. doi: 10.1109/TSMCB.2012.2193613

Dilettoso, E., and Salerno, N. (2006). A self-adaptive niching genetic algorithm for multimodal optimization of electromagnetic devices. *IEEE Trans. Magn.* 42, 1203–1206. doi: 10.1109/TMAG.2006.871672

Dorigo, M. (1992). *Optimization, learning and natural algorithms* (Ph.D. Thesis). Politecnico diMilano, Italy.

Dorigo, M., and Birattari, M. (2010). “Ant colony optimization,” in *Encyclopedia of Machine Learning*, eds C. Sammut and G. I. Webb (Boston, MA: Springer). doi: 10.1007/978-0-387-30164-8

Dorigo, M., and Di Caro, G. (1999). “The ant colony optimization meta-heuristic,” in *New Ideas in Optimization*, eds D. Corne, M. Dorigo, and F. Glover (London: McGraw Hill), 11–32.

Douglas, A. (1951). Laboratory studies of the lambda 4050 group of cometary spectra. *Astrophys. J.* 114:466. doi: 10.1086/145486

Doye, J. P., Wales, D. J., and Miller, M. A. (1998). Thermodynamics and the global optimization of Lennard-Jones clusters. *J. Chem. Phys.* 109, 8143–8153. doi: 10.1063/1.477477

Du, W., and Li, B. (2008). Multi-strategy ensemble particle swarm optimization for dynamic optimization. *Inf. Sci.* 178, 3096–3109. doi: 10.1016/j.ins.2008.01.020

Eberhart, R. C., and Shi, Y. (2000). “Comparing inertia weights and constriction factors in particle swarm optimization,” in *Evolutionary Computation, 2000. Proceedings of the 2000 Congress on: IEEE* (La Jolla, CA: IEEE), 84–88.

Eberhart, R. C., and Shi, Y. (2001). “Tracking and optimizing dynamic systems with particle swarms,” in *Evolutionary Computation, Proceedings of the 2001 Congress on*: *IEEE* (Seoul: IEEE), 94–100.

Engelbrecht, A. P., and Van Loggerenberg, L. (2007). “Evolutionary computation,” in *CEC* (Singapore: IEEE Congress), 2297–2302. doi: 10.1109/CEC.2007.4424757

Fattahi, H., Gholami, A., Amiribakhtiar, M. S., and Moradi, S. (2015). Estimation of asphaltene precipitation from titration data: a hybrid support vector regression with harmony search. *Neural. Comput. Appl.* 26, 789–798. doi: 10.1007/s00521-014-1766-y

Fister, I., and Žumer, J. B. (2012). “Memetic artificial bee colony algorithm for large-scale global optimization,” in *2012 IEEE Congress on Evolutionary Computation* (Brisbane, QLD: IEEE), 1–8.

Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., et al. (2013). *Gaussian 09, Revision D.01*. Wallingford, CT: Gaussian, Inc.

Fulara, J., Lessen, D., Freivogel, P., and Maier, J. (1993). Laboratory evidence for highly unsaturated hydrocarbons as carriers of some of the diffuse interstellar bands. *Nature* 366:439. doi: 10.1038/366439a0

Gavrilas, M., Gavrilas, G., and Sfintes, C. V. (2010). “Application of honey bee mating optimization algorithm to load profile clustering,” in *2010 IEEE*

International Conference on Computational Intelligence for Measurement Systems and Applications (Taranto: IEEE), 113–118.

Geem, Z. W. (2000). *Optimal cost design of water distribution networks using harmony search* (Dissertation). Korea University.

Geem, Z. W. (2006). Optimal cost design of water distribution networks using harmony search. *Eng. Optim.* 38, 259–277. doi: 10.1080/03052150500467430

Geem, Z. W. (ed.) (2001). *Music-Inspired Harmony Search Algorithm*. Berlin: Springer.

Geem, Z. W., Kim, J. H., and Loganathan, G. V. (2001). A new heuristic optimization algorithm: harmony search. *Simulation* 76, 60–68. doi: 10.1177/003754970107600201

Geem, Z. W., Lee, K. S., and Park, Y. (2005). Application of harmony search to vehicle routing. *Am. J. Appl. Sci.* 2, 1552–1557. doi: 10.3844/ajassp.2005.1552.1557

Gholizadeh, S., and Barzegar, A. (2013). Shape optimization of structures for frequency constraints by sequential harmony search algorithm. *Eng. Optim.* 45, 627–646. doi: 10.1080/0305215X.2012.704028

Glass, C. W., Oganov, A. R., and Hansen, N. (2006). USPEX—Evolutionary crystal structure prediction. *Comput. Phys. Commun.* 175, 713–720. doi: 10.1016/j.cpc.2006.07.020

Goh, C. K., Tan, K. C., Liu, D., and Chiam, S. C. (2010). A competitive and cooperative co-evolutionary approach to multi-objective particle swarm optimization algorithm design. *Eur. J. Operat. Res.* 202, 42–54. doi: 10.1016/j.ejor.2009.05.005

Grüninger, T., and Wallace, D. (1996). *Multimodal optimization using genetic algorithms* (Master's Thesis). Stuttgart University.

Guangneng, F., Lixia, H., and Xueguang, H. (2005). Synthesis of single-crystal BaTiO3 nanoparticles via a one-step sol-precipitation route. *J. Cryst. Growth* 279, 489–493. doi: 10.1016/j.jcrysgro.2005.02.054

Haddad, O. B., Afshar, A., and Mariño, M. A. (2006). Honey-bees mating optimization (HBMO) algorithm: a new heuristic approach for water resources optimization. *Water Res. Manag.* 20, 661–680. doi: 10.1007/s11269-005-9001-3

Hadwan, M., Ayob, M., Sabar, N. R., and Qu, R. (2013). A harmony search algorithm for nurse rostering problems. *Inform. Sci.* 233, 126–140. doi: 10.1016/j.ins.2012.12.025

Hassan, R., Cohanim, B., de Weck, O., and Venter, G. (2005). "A comparison of particle swarm optimization and the genetic algorithm," in *Proceedings of the 46thAIAA/ASME/ASCE/AHS/ASC Structures, Structural Dy-namics and Materials Conference* (Austin, TX).

Heppner, F., and Grenander, U. (1990). "A stochastic nonlinear model for coordinated bird flocks," in *The Ubiquity of Chaos*, eds S. Krasner (AAAS Publications), 233–238.

Hoang, D. C., Yadav, P., Kumar, R., and Panda, S. K. (2014). Real-time implementation of a harmony search algorithm-based clustering protocol for energy-efficient wireless sensor networks. *IEEE Trans. Industr. Inform.* 10, 774–783. doi: 10.1109/TII.2013.2273739

Holland, J. H. (1992). *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. MIT press.

Hutter, J., Luethi, H. P., and Diederich, F. (1994). Structures and vibrational frequencies of the carbon molecules C2-C18 calculated by density functional theory. *J. Am. Chem. Soc.* 116, 750–756. doi: 10.1021/ja00081a041

Hutter, J., and Lüthi, H. P. (1994). The molecular structure of C_6: a theoretical investigation. *J. Chem. Phys.* 101, 2213–2216. doi: 10.1063/1.467661

Inderscience (2010). *Cuckoo Designs Spring*. Retrieved from: Alphagalileo.org

Jahanshahi, G., and Haddad, O. B. (2008). "Honey-bee mating optimization (HBMO) algorithm for optimal design of water distribution systems," in *World Environmental and Water Resources Congress 2008: Ahupua'A*, (Honolulu, HI). 1–16.

Karaboga, D., and Basturk, B. (2007). A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. *J. Glob. Optim.* 39, 459–471. doi: 10.1007/s10898-007-9149-x

Karaboga, D., and Basturk, B. (2008). On the performance of artificial bee colony (ABC) algorithm. *Appl. Soft Comput.* 8, 687–697. doi: 10.1016/j.asoc.2007.05.007

Kennedy, J. (1997). "The particle swarm: social adaptation of knowledge," in *Proceedings of 1997 IEEE International Conference on Evolutionary Computation (ICEC'97)* (Indianapolis, IN: IEEE), 303–308.

Kennedy, J., and Eberhart, R. (1995). "Particle swarm optimization (PSO)", in: *Proceedings of IEEE International Conference on Neural Networks* (Perth, WA: IEEE), 1942–1948.

Kennedy, J., and Eberhart, R. C. (1999). "The particle swarm: social adaptation in information-processing systems," in *New Ideas in Optimization* (McGraw-Hill Ltd.), 379–388.

Khan, A., and Sadeequllah, M. (2010). "Rank based particle swarm optimization," in *International Conference on Swarm Intelligence* (Berlin; Heidelberg: Springer), 275–286. doi: 10.1007/978-3-642-15461-4_24

Kiranyaz, S., Pulkkinen, J., and Gabbouj, M. (2011). "Multi-dimensional PSO for dynamic environments," in *International Conference on Innovations in Information Technology* (Tampere), 2212–2223. doi: 10.1016/j.eswa.2010.08.009

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* 220, 671–680. doi: 10.1126/science.220.4598.671

Koinuma, H., Horiuchi, T., Inomata, K., Ha, H.-K., Nakajima, K., and Chaudhary, K. (1996). Synthesis of carbon clusters and thin films by low temperature plasma chemical vapor deposition under atmospheric pressure. *Pure Appl. Chem.* 68, 1151–1154. doi: 10.1351/pac199668051151

Korošec, P., Šilc, J., and Filipič, B. (2012). The differential ant-stigmergy algorithm. *Inform. Sci.* 192, 82–97. doi: 10.1016/j.ins.2010.05.002

Kroto, H., and McKay, K. (1988). The formation of quasi-icosahedral spiral shell carbon particles. *Nature* 331:328. doi: 10.1038/331328a0

Krug, M., Nguang, S. K., Wu, J., and Shen, J. (2010). GA-based model predictive control of boiler-turbine systems. *Int. J. Innov. Comput. Inf. Control* 6, 5237–5248. Available online at: http://www.ijicic.org/09-0646-1.pdf

Lee, C., Yang, W., and Parr, R. G. (1988). Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* 37:785. doi: 10.1103/PhysRevB.37.785

Li, G., Niu, P., and Xiao, X. (2012). Development and investigation of efficient artificial bee colony algorithm for numerical function optimization. *Appl. Soft Comput.* 12, 320–332. doi: 10.1016/j.asoc.2011.08.040

Li, H.-Q., and Li, L. (2007). "A novel hybrid particle swarm optimization algorithm combined with harmony search for high dimensional optimization problems," in *Intelligent Pervasive Computing, IPC. The International Conference on: IEEE* (Jeju City: IEEE), 94–97.

Li, M., Lin, D., and Kou, J. (2012). A hybrid niching PSO enhanced with recombination-replacement crowding strategy for multimodal function optimization. *Appl. Soft Comput.* 12, 975–987. doi: 10.1016/j.asoc.2011.11.032

Li, X. (2007). "A multimodal particle swarm optimizer based on fitness Euclidean-distance ratio," in *Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation: ACM* (New York, NY), 78–85. doi: 10.1145/1276958.1276970

Li-Ping, Z., Huan-Jun, Y., and Shang-Xu, H. (2005). Optimal choice of parameters for particle swarm optimization. *J. Zhejiang Univ. Sci. A* 6, 528–534. doi: 10.1631/jzus.2005.A0528

Liu, B., Wang, L., and Jin, Y.-H. (2007). An effective PSO-based memetic algorithm for flow shop scheduling. *IEEE Trans. Syst. Man Cybern. B* 37, 18–27. doi: 10.1109/TSMCB.2006.883272

Liu, D., Tan, K. C., Goh, C. K., and Ho, W. K. (2007). A multiobjective memetic algorithm based on particle swarm optimization. *IEEE Trans. Syst. Man Cybern. B* 37, 42–50. doi: 10.1109/TSMCB.2006.883270

Liu, X., Liu, H., and Duan, H. (2007). Particle swarm optimization based on dynamic niche technology with applications to conceptual design. *Adv. Eng. Softw.* 38, 668–676. doi: 10.1016/j.advengsoft.2006.10.009

Liu, Z., and Wang, S. (2006). "Hybrid particle swarm optimization for permutation flow shop scheduling," in *Intelligent Control and Automation, WCICA 2006. The Sixth World Congress on: IEEE* (Dalian: IEEE), 3245–3249.

Ma, Q., Lei, X., and Zhang, Q. (2009). "Mobile robot path planning with complex constraints based on the second-order oscillating particle swarm optimization algorithm," in *Computer Science and Information Engineering, WRI World Congress on*: *IEEE* (Los Angeles, CA: IEEE), 244–248.

Manjarres, D., Landa-Torres, I., Gil-Lopez, S., Del Ser, J., Bilbao, M. N., Salcedo-Sanz, S., et al. (2013). A survey on applications of the harmony search algorithm. *Eng. Appl. Artif. Intell.* 26, 1818–1831. doi: 10.1016/j.engappai.2013.05.008

Marinaki, M., Marinakis, Y., and Zopounidis, C. (2010). Honey bees mating optimization algorithm for financial classification problems. *Appl. Soft Comput.* 10, 806–812. doi: 10.1016/j.asoc.2009.09.010

Marinakis, Y., and Marinaki, M. (2009). "A hybrid honey bees mating optimization algorithm for the probabilistic traveling salesman problem," in *2009 IEEE Congress on Evolutionary Computation: IEEE* (Trondheim: IEEE), 1762–1769.

Martin, J., François, J.-P., and Gijbels, R. (1993). The impact of quantum chemical methods on the interpretation of molecular spectra of carbon clusters. *J. Mol. Struct.* 294, 21–24. doi: 10.1016/0022-2860(93)80305-F

Martin, J. M., and Taylor, P. R. (1996). Structure and vibrations of small carbon clusters from coupled-cluster calculations. *J. Phys. Chem.* 100, 6047–6056. doi: 10.1021/jp952471r

Martonák, R., Laio, A., Bernasconi, M., Ceriani, C., Raiteri, P., Zipoli, F., et al. (2005). Simulation of structural phase transitions by metadynamics. *Z. Kristallogr. Cryst. Mater.* 220, 489–498. doi: 10.1524/zkri.220.5.489.65078

Martonák, R., Laio, A., and Parrinello, M. (2003). Predicting crystal structures: the Parrinello-Rahman method revisited. *Phys. Rev. Lett.* 9:075503. doi: 10.1103/PhysRevLett.90.075503

Millonas, M. M. (1993). *Swarms, Phase Transitions, and Collective Intelligence (Paper 1); and a Nonequilibrium Statistical Field Theory of Swarms and Other Spatially Extended Complex Systems (Paper 2)* (No. 93-06-039).

Mitikiri, P., Jana, G., Sural, S., and Chattaraj, P. K. (2018). A machine learning technique toward generating minimum energy structures of small boron clusters. *Int. J. Quantum Chem.* 118:e25672. doi: 10.1002/qua.25672

Mujica, A., and Needs, R. (1997). Erratum: theoretical study of the high-pressure phase stability of GaP, InP, and InAs. *Phys. Rev. B* 56:12653. doi: 10.1103/PhysRevB.56.12653

Nasrinpour, H., Bavani, A., and Teshnehlab, M. (2017). Grouped bees algorithm: a grouped version of the bees algorithm. *Computers* 6:5. doi: 10.3390/computers6010005

Nayeem, A., Vila, J., and Scheraga, H. A. (1991). A comparative study of the simulated-annealing and Monte Carlo-with-minimization approaches to the minimum-energy structures of polypeptides:[Met]-enkephalin. *J. Comput. Chem.*12, 594–605. doi: 10.1002/jcc.540120509

Nekooei, K., Farsangi, M. M., Nezamabadi-Pour, H., and Lee, K. Y. (2013). An improved multi-objective harmony search for optimal placement of DGs in distribution systems. *IEEE Trans. Smart Grid* 4, 557–567. doi: 10.1109/TSG.2012.2237420

Nickabadi, A., Ebadzadeh, M. M., and Safabakhsh, R. (2008). "DNPSO: A dynamic niching particle swarm optimizer for multi-modal optimization," in *Evolutionary Computation, 2008. CEC 2008* (Hong Kong: IEEE), 26–32.

Oganov, A. R., and Glass, C. W. (2006). Crystal structure prediction using ab initio evolutionary techniques: principles and applications. *J. Chem. Phys.* 124:244704. doi: 10.1063/1.2210932

Omkar, S., Senthilnath, J., Khandelwal, R., Naik, G. N., and Gopalakrishnan, S. (2011). Artificial Bee Colony (ABC) for multi-objective design optimization of composite structures. *Appl. Soft Comput.* 11, 489–499. doi: 10.1016/j.asoc.2009.12.008

Özcan, E., and Yilmaz, M. (2007). "Particle swarms for multimodal optimization," in *International Conference on Adaptive and Natural Computing Algorithms* (Berlin; Heidelberg: Springer), 366–375.

Pannetier, J., Bassas-Alsina, J., Rodriguez-Carvajal, J., and Caignaert, V. (1990). Prediction of crystal structures from crystal chemistry rules by simulated annealing. *Nature* 346, 343–345. doi: 10.1038/346343a0

Payne, R. B., and Sorensen, M. D. (2005). *The Cuckoos.* New York, NY: Oxford University Press.

Pedersen, M. E. H. (2010). *Good Parameters for Particle Swarm Optimization*. Technical Report HL1001, Hvass Lab, Copenhagen.

Petalas, Y. G., Parsopoulos, K. E., Papageorgiou, E. I., Groumpos, P. P., and Vrahatis, M. N. (2007). "Enhanced learning in fuzzy simulation models using memetic particle swarm optimization," in *Swarm Intelligence Symposium, SIS, IEEE* (Honolulu, HI: IEEE), 16–22.

Pham, D. T., and Castellani, M. (2009). The bees algorithm: modelling foraging behaviour to solve continuous optimization problems. *Proc. Inst. Mech. Eng. C* 223, 2919–2938. doi: 10.1243/09544062JMES1494

Pham, D. T., and Castellani, M. (2014). Benchmarking and comparison of nature-inspired population-based continuous optimisation algorithms. *Soft. Comput.* 18, 871–903. doi: 10.1007/s00500-013-1104-9

Pham, D. T., and Castellani, M. (2015). A comparative study of the Bees Algorithm as a tool for function optimisation. *Cogent Eng.* 2:1091540. doi: 10.1080/23311916.2015.1091540

Pham, D. T., Ghanbarzadeh, A., Koc, E., Otri, S., Rahim, S., and Zaidi, M. (2005). *The Bees Algorithm*. Technical Note, Manufacturing Engineering Centre, Cardiff University, UK.

Pickard, C. J., and Needs, R. (2006). High-pressure phases of silane. *Phys. Rev. Lett.* 97:045504. doi: 10.1103/PhysRevLett.97.045504

Pickard, C. J., and Needs, R. (2008). Highly compressed ammonia forms an ionic crystal. *Nat. Mat.* 7, 775–779. doi: 10.1038/nmat2261

Pickard, C. J., and Needs, R. J. (2007). Structure of phase III of solid hydrogen. *Nat. Phys.* 3, 473–476. doi: 10.1038/nphys625

Pitzer, K. S., and Clementi, E. (1959). Large molecules in carbon vapor. *J. Am. Chem. Soc.* 81, 4477–4485. doi: 10.1021/ja01526a010

Pless, V., Suter, H., and Engels, B. (1994). Ab initio study of the energy difference between the cyclic and linear forms of the C_6 molecule. *J. Chem. Phys.*101, 4042–4048. doi: 10.1063/1.467521

Poli, R. (2007). *An Analysis of Publications on Particle Swarm Optimization Applications*. Essex: Department of Computer Science, University of Essex.

Poli, R. (2008). Analysis of the publications on the applications of particle swarm optimisation. *J. Artif. Evol. Appl.* 2008:685175. doi: 10.1155/2008/685175

Poli, R., and Langdon, W. B. (2002). *Foundations of Genetic Programming.* Berlin: Springer.

Price, K., Storn, R. M., and Lampinen, J. A. (2006). *Differential Evolution: A Practical Approach to Global Optimization*. Berlin; Heidelberg: Springer. doi: 10.1007/3-540-31306-0

Qu, B.-Y., Liang, J. J., and Suganthan, P. N. (2012). Niching particle swarm optimization with local search for multi-modal optimization. *Inform. Sci.* 197, 131–143. doi: 10.1016/j.ins.2012.02.011

Raghavachari, K., and Binkley, J. (1987). Structure, stability, and fragmentation of small carbon clusters. *J. Chem. Phys.* 87, 2191–2197. doi: 10.1063/1.453145

Rajasekhar, A., Lynn, N., Das, S., and Suganthan, P. N. (2017). Computing with the collective intelligence of honey bees–a survey. *Swarm Evol. Comput.* 32, 25–48. doi: 10.1016/j.swevo.2016.06.001

Reeves, W. T. (1983). Particle systems—a technique for modeling a class of fuzzy objects. *ACM Trans. Graph.* 2, 91–108. doi: 10.1145/357318.357320

Reynolds, C. W. (1987). Flocks, herds and schools: a distributed behavioral model. *ACM SIGGRAPH Comput. Graph.* 21, 25–34. doi: 10.1145/37402.37406

Richardson, P. (2008). *Bats.* London: Natural History Museum.

Rocca, P., Oliveri, G., and Massa, A. (2011). Differential evolution as applied to electromagnetics. *IEEE Antennas Propag. Mag.* 53, 38–49. doi: 10.1109/MAP.2011.5773566

Schutze, O., Talbi, E.-G., Pulido, G. T., Coello, C. C., and Santana-Quintero, L. V. (2007). "A memetic PSO algorithm for scalar optimization problems," in *Swarm Intelligence Symposium*, *SIS* 2007 (Washington, DC: IEEE Computer Society), 128–134.

Shao, X., Cheng, L., and Cai, W. (2004). A dynamic lattice searching method for fast optimization of Lennard–Jones clusters. *J. Comput. Chem.* 25, 1693–1698. doi: 10.1002/jcc.20096

Shao, X., Yang, X., and Cai, W. (2008). A dynamic lattice searching method with interior operation for unbiased optimization of large Lennard-Jones clusters. *J. Comput. Chem.* 29, 1772–1779. doi: 10.1002/jcc.20938

Shi, Y. (2001). "Particle swarm optimization: developments, applications and resources," in *Evolutionary Computation, Proceedings of the 2001 Congress on*: *IEEE* (Seoul: IEEE), 81–86.

Shi, Y., and Eberhart, R. (1998). "A modified particle swarm optimizer," in *Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence. IEEE International Conference* (Anchorage, AK: IEEE), 69–73.

Sivasubramani, S., and Swarup, K. S. (2009). "Multiagent based particle swarm optimization approach to economic dispatch with security constraints," in *Power Systems, ICPS'09. International Conference on: IEEE* (Kharagpur: IEEE), 1–6.

Storn, R. (1996). "On the usage of differential evolution for function optimization," in *Proceedings of North American Fuzzy Information Processing* (Berkeley, CA: IEEE), 519–523.

Storn, R., and Price, K. (1997). Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *J. Glob. Optim.* 11, 341–359. doi: 10.1023/A:1008202821328

Sun, C., Liang, H., Li, L., and Liu, D. (2007). "Clustering with a weighted sum validity function using a niching PSO algorithm," in *Networking, Sensing and Control, 2007 IEEE International Conference on*: *IEEE* (London: IEEE), 368–373.

Sun, L.-Q., and Gao, X.-Y. (2008). "Improved chaos-particle swarm optimization algorithm for geometric constraint solving," in *Computer Science and Software Engineering, International Conference on*: *IEEE* (Hubei: IEEE), 992–995.

Talbi, E.-G. (2009). *Metaheuristics: From Design to Implementation*. Hoboken, NJ: John Wiley & Sons, Inc.

Trelea, I. C. (2003). The particle swarm optimization algorithm: convergence analysis and parameter selection. *Inform. Proc. Lett.* 85, 317–325. doi: 10.1016/S0020-0190(02)00447-7

Trimarchi, G., and Zunger, A. (2007). Global space-group optimization problem: Finding the stablest crystal structure without constraints. *Phys. Rev. B* 75, 104113. doi: 10.1103/PhysRevB.75.104113

Unler, A., and Murat, A. (2010). A discrete particle swarm optimization method for feature selection in binary classification problems. *Eur. J. Oper. Res.* 206, 528–539. doi: 10.1016/j.ejor.2010.02.032

Ursem, R. K. (2000). "Multinational GAs: multimodal optimization techniques in dynamic environments," in *GECCO*, Las Vegas, NV; San Francisco, CA: Morgan Kaufmann Publishers Inc, 19–26.

Van Orden, A., and Saykally, R. J. (1998). Small carbon clusters: spectroscopy, structure, and energetics. *Chem Rev.* 98, 2313–2358. doi: 10.1021/cr970086n

Wales, D. J., and Doye, J. P. (1997). Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *J. Phys. Chem. A* 101, 5111–5116. doi: 10.1021/jp970984n

Wang, J., Liu, D., and Shang, H. (2009). "Artificial intelligence and computational intelligence," in *AICI'09. International Conference* (Shanghai), 139–144.

Wang, L., and Li, L.-P. (2013). An effective differential harmony search algorithm for the solving non-convex economic load dispatch problems. *Int. J. Elec. Power* 44, 832–843. doi: 10.1016/j.ijepes.2012.08.021

Wang, Y., Chen, P., and Jin, Y. (2009). "Trajectory planning for an unmanned ground vehicle group using augmented particle swarm optimization in a dynamic environment," in *Systems, Man and Cybernetics, SMC. IEEE International Conference on: IEEE* (San Antonio, TX: IEEE), 4341–4346.

Wang, Y., Li, B., Weise, T., Wang, J., Yuan, B., and Tian, Q. (2011). Self-adaptive learning based particle swarm optimization. *Inform. Sci.* 181, 4515–4538. doi: 10.1016/j.ins.2010.07.013

Wang, Z., and Xing, H. (2008). "Dynamic-probabilistic particle swarm synergetic model: A new framework for a more in-depth understanding of particle swarm algorithms," in *Evolutionary Computation, CEC 2008. (IEEE World Congress on Computational Intelligence). IEEE Congress on: IEEE* (Hong Kong: IEEE), 312–321.

Watts, J. D., Gauss, J., Stanton, J. F., and Bartlett, R. J. (1992). Linear and cyclic isomers of C4. A theoretical study with coupled-cluster methods and large basis sets. *J. Chem. Phys.* 97, 8372–8381. doi: 10.1063/1.463407

Weltner, W. Jr., and Van Zee, R. J. (1989). Carbon molecules, ions, and clusters. *Chem. Rev.* 89, 1713–1747. doi: 10.1021/cr00098a005

Weyland, D. (2015). A critical analysis of the harmony search algorithm—How not to solve sudoku. *Oper. Res. Persp.* 2, 97–105. doi: 10.1016/j.orp.2015.04.001

Woodley, S., Battle, P., Gale, J., and Catlow, C. A. (1999). The prediction of inorganic crystal structures using a genetic algorithm and energy minimisation. *Phys. Chem. Chem. Phys.* 1, 2535–2542. doi: 10.1039/a901227c

Yang, X., Yuan, J., Yuan, J., and Mao, H. (2007). A modified particle swarm optimizer with dynamic adaptation. *Appl. Mat. Comput.* 189, 1205–1213. doi: 10.1016/j.amc.2006.12.045

Yang, X.-S. (2010a). *Nature-Inspired Metaheuristic Algorithms, 2nd Edn*. Cambridge, UK: University of Cambridge; Luniver Press.

Yang, X.-S. (2010b). "A new metaheuristic bat-inspired algorithm," in *Nature Inspired Cooperative Strategies for Optimization (NICSO 2010)* (Berlin; Heidelberg: Springer), 65–74.

Yang, X.-S., and Deb, S. (2009). "Cuckoo search via Lévy flights," in *World Congress on Nature & Biologically Inspired Computing (NaBIC)* (Coimbatore: IEEE), 210–214.

Yeh, W.-C. (2009). A two-stage discrete particle swarm optimization for the problem of multiple multi-level redundancy allocation in series systems. *Expert Sys. Appl.* 36, 9192–9200. doi: 10.1016/j.eswa.2008.12.024

Yeh, W.-C., Chang, W.-W., and Chung, Y. Y. (2009). A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method. *Expert Sys. Appl.* 36, 8204–8211. doi: 10.1016/j.eswa.2008.10.004

Yin, P.-Y. (2004). A discrete particle swarm algorithm for optimal polygonal approximation of digital curves. *J. Vis. Commun. Image Represent.* 15, 241–260. doi: 10.1016/j.jvcir.2003.12.001

Zhan, Z.-H., Zhang, J., Li, Y., and Chung, H. S. (2009). Adaptive particle swarm optimization. *IEEE Trans. Syst. Man Cybern. B* 39, 1362–1381. doi: 10.1109/TSMCB.2009.2015956

Zhang, J., Huang, D.-S., and Liu, K.-H. (2007). "Multi-sub-swarm particle swarm optimization algorithm for multimodal function optimization," in *Evolutionary Computation, CEC, IEEE Congress on: IEEE* (Singapore: IEEE), 3215–3220.

Zhang, J., Xie, L., and Wang, S. (2006). Particle swarm for the dynamic optimization of biochemical processes. *Comp. Aided Chem. Eng.* 21, 497–502. doi: 10.1016/S1570-7946(06)80094-5

Zhang, R., and Wang, D. (2008). "Forecasting annual electricity demand using BP neural network based on three sub-swarms PSO," in *Control and Decision Conference, CCDC 2008* (Yantai: IEEE), 1409–1413.

Zhao, S.-Z., Liang, J. J., Suganthan, P. N., and Tasgetiren, M. F. (2008). "Dynamic multi-swarm particle swarm optimizer with local search for large scale global optimization," in *Evolutionary Computation, CEC. (IEEE World Congress on Computational Intelligence). IEEE Congress on: IEEE* (Hong Kong: IEEE), 3845–3852.

Zheng, S.-F., Hu, S.-L., Su, S.-X., Lin, C.-F., and Lai, X.-W. (2007). "A modified particle swarm optimization algorithm and application," in *International Conference on Machine Learning and Cybernetics* (Guangzhou: IEEE), 945–951.

Zhi-Jie, L., Xiang-Dong, L., Xiao-Dong, D., and Cun-Rui, W. (2009). "An improved particle swarm algorithm for search optimization," in *WRI Global Congress on Intelligent Systems* (Xiamen: IEEE), 154–158.

Zlochin, M., Birattari, M., Meuleau, N., and Dorigo, M. (2004). Model-based search for combinatorial optimization: a critical survey. *Annal. Oper. Res.* 131, 373–395. doi: 10.1023/B:ANOR.0000039526.52305.af

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Chemistry

REVIEW
published: 07 August 2019
doi: 10.3389/fchem.2019.00519

Augmenting Basin-Hopping With Techniques From Unsupervised Machine Learning: Applications in Spectroscopy and Ion Mobility

*Ce Zhou, Christian Ieritano and William Scott Hopkins**

Department of Chemistry, University of Waterloo, Waterloo, ON, Canada

Evolutionary algorithms such as the basin-hopping (BH) algorithm have proven to be useful for difficult non-linear optimization problems with multiple modalities and variables. Applications of these algorithms range from characterization of molecular states in statistical physics and molecular biology to geometric packing problems. A key feature of BH is the fact that one can generate a coarse-grained mapping of a potential energy surface (PES) in terms of local minima. These results can then be utilized to gain insights into molecular dynamics and thermodynamic properties. Here we describe how one can employ concepts from unsupervised machine learning to augment BH PES searches to more efficiently identify local minima and the transition states connecting them. Specifically, we introduce the concepts of similarity indices, hierarchical clustering, and multidimensional scaling to the BH methodology. These same machine learning techniques can be used as tools for interpreting and rationalizing experimental results from spectroscopic and ion mobility investigations (e.g., spectral assignment, dynamic collision cross sections). We exemplify this in two case studies: (1) assigning the infrared multiple photon dissociation spectrum of the protonated serine dimer and (2) determining the temperature-dependent collision cross-section of protonated alanine tripeptide.

Keywords: serine dimer, polyalanine, collision cross section, IRMPD, hierarchical clustering, potential energy surface, global optimization, vibrational spectroscopy

OPEN ACCESS

Edited by:
Emilio Martinez-Nunez,
University of Santiago de Compostela, Spain

Reviewed by:
Daniel Pelaez,
Université de Lille, France
Tomás González-Lezana,
Spanish National Research Council (CSIC), Spain

****Correspondence:***
William Scott Hopkins
scott.hopkins@uwaterloo.ca

Specialty section:
This article was submitted to Theoretical and Computational Chemistry,
a section of the journal Frontiers in Chemistry

Received: *30 April 2019*
Accepted: *08 July 2019*
Published: *07 August 2019*

Citation:
Zhou C, Ieritano C and Hopkins WS (2019) Augmenting Basin-Hopping With Techniques From Unsupervised Machine Learning: Applications in Spectroscopy and Ion Mobility. Front. Chem. 7:519. doi: 10.3389/fchem.2019.00519

INTRODUCTION

Molecular global optimization (GO) to identify the chemically-relevant species on hypergeometric potential energy surfaces (PESs) provides both rationalizations and predictions of experimental observations by relating thermodynamic and kinetic properties to the accessible local minima and the transition states (TSs) that connect them (Scheraga, 1992; Piela et al., 1994; Wales and Doye, 1997; Wales and Scheraga, 1999). Basin-hopping (BH) is a technique for GO that is based on the iterative approach of performing random perturbation of geometric coordinates, local optimization of a model potential energy function, and accepting or rejecting the perturbed coordinates based on the value of the minimized function (Wales and Doye, 1997; Wales et al., 1998; Wales and Scheraga, 1999; Lecours et al., 2014). Use of the BH algorithm for searching molecular PESs was outlined by Wales and Doye in their 1997 article "Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms," (Wales and Doye, 1997) which describes how the technique transforms the PES into a collection of interpenetrating staircases wherein each stair/plateau on the transformed surface is associated with a stationary point (usually local minimum) of the original potential energy landscape. **Figure 1** shows a flow diagram outlining the general procedure of the BH search algorithm. The key feature of the BH algorithm is the inclusion of assessment criteria for accepting or rejecting a newly distorted input geometry. One of these criteria is the definite replacement of the lowest energy structure identified by the BH routine with the currently optimized structure if that structure has a lower energy. A second key criterion is a conditional acceptance of the distorted geometry by assessing the statistical accessibility of the optimized structure based on a pre-defined energy window. For example, one can define a Boltzmann distribution at a given temperature with respect to the current lowest energy structure and assess the probability of accessing the newly generated stationary point. Thus, the BH algorithm has a bias toward low energy structures and is a good option for identifying the global minimum (GM) and local minima that may be present in an ensemble under thermal equilibrium conditions.

To further improve the efficiency of a BH search, one can include additional criteria for assessment of distorted molecular geometries prior to optimization. For example, one might choose to reject structures in which inter-atomic distances are less than some pre-defined threshold, or one might choose to define an interaction volume to prevent molecular/cluster dissociation (Lecours et al., 2014). It is also common to select specific degrees of freedom (DoFs) for random distortion while freezing others; one might choose to search the conformational space defined by molecular dihedral angles while leaving the distances between chemically bonded atoms fixed (Hopkins et al., 2013, 2015). There are several other works which employ more dramatic changes to the underlying BH algorithm. For example, Leary proposed a version in which only the replacement criterion is employed in the evaluation (i.e., no statistically accessible energy window is specified) (Leary, 2000). In other works, Röder and Wales propose a mutational BH algorithm to optimize biomolecules (Röder and Wales, 2018), and Kim et al. combine BH with Coulomb matrix analysis to sample reaction intermediates (Kim et al., 2014). While these variants have all been successful in the task at hand, the fact that the basic BH algorithm often requires tailoring highlights the inherent drawbacks in the BH methodology.

One principal short-coming of the BH algorithm that practitioners must be aware of is that the method is not deterministic; i.e., identifying the GM via a finite, stochastic search is not guaranteed. Confidence in BH search results come from a satisfactory agreement with experimental observations and/or the consistency of results from several parallel simulations with different initial conditions. A second potential short-coming is the fact that, due to performance considerations, BH calculations are often conducted with relatively low-level model chemistries (e.g., molecular mechanics), which may not be accurate enough for certain molecular systems. Finally, practitioners must be aware that a BH search may be kinetically trapped in a local potential minimum if the thermal energy (*viz.* temperature) of the simulation is set too low. In fact, in some cases BH searches of PESs are non-ergodic regardless of simulation temperature. For example, consider the case of protonated *para*-aminobenzoic acid, which can exhibit protonation on either the carbonyl oxygen atom or the amine nitrogen atom in the gas phase (Tian and Kass, 2009; Schmidt et al., 2011; Campbell et al., 2012, 2016). If one were to assume that the protonation site of *para*-aminobenzoic acid were the nitrogen center (as is the case in protic solution) and model the system as a molecular cation using a molecular mechanics force field, the O-protonated isomer (which is the gas phase global minimum) would not be identified without modifying the atomic connectivity during the BH search (Tian and Kass, 2008; Campbell et al., 2012, 2016). To overcome this systematic limitation, one must treat the charge-carrying proton as a separate moiety in the simulation and/or augment the BH framework with the chemical intuition of the user (i.e., manually identify both prototropic isomers and conduct BH searches for each of them).

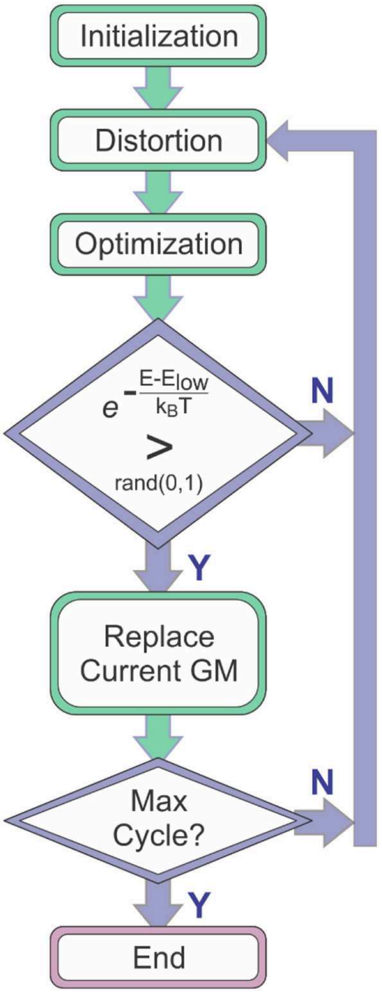


FIGURE 1 | The general procedure of the basin-hopping algorithm. E_{low} is the energy of the lowest energy species identified to that point in the search (i.e., the current global minimum, GM).

Here, we describe how the basin-hopping algorithm can be employed to reliably model gas phase cluster and molecular systems for comparison with observations from spectroscopy and ion mobility experiments. To model our experimental observations, we require theoretical predictions from a collection of local minima, which do not necessarily include the global minimum, and an efficient method to find matches between the predictions and the observations. In some cases, it is also desirable to identify the TSs that connect minima to assess thermodynamic accessibility of the various isomers / conformers. These two requirements present two notable challenges for the BH methodology. The first challenge, related to the principal short-coming mentioned above, is the necessity to accurately track the explored regions of the PES. In doing so, one not only identifies a set of local minima, but also gains useful information for directing the BH search toward regions of the PES that are relatively unexplored. The second challenge is the accurate and efficient identification of the TSs that connect local minima. To overcome these challenges, we collect the nuclear configuration data that is generated during the BH search and utilize this data as described in Section Augmenting the BH Algorithm. Specifically, in Section Assessing Geometric Similarity we describe how one can utilize similarity functions and hierarchical clustering, which are concepts generally associated with unsupervised machine learning, to assess the uniqueness of the local minima and guide PES searches. We then discuss the interpolation of geometries to identify intermediate local minima and to create guess geometries for TS searches in Section Interpolating Intermediate Geometries. In Section Application of BH Search Results, we outline our methods for employing our BH results to assign the spectral carriers (Section Case Study 1: The IR Spectrum of the Protonated Serine Dimer) and to model temperature-dependent structures (Section Case Study 2: Dynamic Collision Cross Section of Protonated Alanine Tripeptide) of geometrically-fluxional species. Finally, we summarize our perspective and highlight open questions in Section Conclusions.

AUGMENTING THE BH ALGORITHM

As mentioned in section Introduction, several variations to the BH algorithm have been proposed to address specific challenges in searching complex potential energy landscapes (Leary, 2000; Kim et al., 2014; Röder and Wales, 2018). For our purposes, where it is necessary to identify a collection of local minima that are representative of the species present in experimental ensembles, we require a faithful mapping of the molecular PES. To improve the efficiency and PES coverage of the BH algorithm, we introduce a method of comparing the geometries of local minima. This comparison, which is derived from a similarity function, provides a more rigorous identification of unique isomeric species and insight into which regions of the PES may require additional exploration.

In analogy to the spatial distance between two locations on a map, a similarity function quantifies the similarity of two conformations, A and B, in conformation space. The function, usually denoted as d(A,B), is non-negative ($d\left(A,B\right)\ \geq\ 0$), symmetric ($d\left(A,B\right)\ =\ d(B,A)$) and has zero value only when two identical elements are evaluated ($d\left(A,A\right)\ =\ 0$) (Locatelli and Schoen, 2013). The similarity function can be used in one of three ways: qualification, quantification, and interpolation. Qualification usage implies that the function need only tell if two input structures are identical. Quantification usage provides a metric for how much difference is there between two structures; for example, is structure A more similar to structure B than to structure C? Interpolation usage means that, given two structures, A and B, and an arbitrary interpolation factor, $\lambda\ \in (0, 1)$, there exist one or more structures, C, satisfying:

$$d\left(A,B\right)=\frac{1}{\lambda}d\left(A,C\right)=\frac{1}{1-\lambda}d\left(B,C\right) \tag{1}$$

If the function d satisfies triangular inequality $d\left(A,B\right)+d\left(B,C\right)\ \geq\ d(A,C)$), the structure C is unique, and d is a metric of the conformation space (Choudhary, 2003). Note that special treatment is required if A and B have different numbers of atoms (i.e., if A and B are of different dimension); this tends not to be the case in simulations of chemical systems. The interpolation mechanism is of central importance not only to a number of GO algorithms, such as particle swarm optimization (Eberhart and Yuhui, 2001), differential evolution (Storn and Price, 1997), and DIRECT (Jones et al., 1993), but also to unsupervised machine learning techniques such as the self-organizing map (Kohonen, 1990) and the growing neural gas (Martinez and Schulten, 1991; Fritzke, 1994). In qualitative comparisons, the similarity function need only account for the translational, rotational, and permutational invariance under a given molecular representation; structural equivalence only occurs between species of identical composition. Such invariance properties are either embedded in the mathematical definition of the molecular representation or they are achieved via manually aligning the two molecular systems prior to evaluating their similarity. Examples of such representations include the conventional skeletal chemical formula and the SMILEs code used in compound database systems (Weininger, 1988; Rahman et al., 2009; Heller et al., 2013). In quantitative comparisons, the similarity of two structures is specified by a real number. These similarity indices are useful in discriminating visited regions of the PES (e.g., well-sampled vs. poorly-sampled regions), which can be assessed using unsupervised machine learning analyses like hierarchical clustering and multidimensional scaling (MDS) (Wickelmaier, 2003; Borg and Groenen, 2005). Most similarity functions used for quantitation purposes are defined by the normal (e.g., the root-mean-square deviation of atomic positions, RMSD) (Kabsch, 1976) or reciprocal (e.g., the Coulomb matrix) (Montavon et al., 2012) interatomic distances, although electron density-based similarity functions have found use in drug discovery (Cereto, 2015; Kumar and Zhang, 2018). To implement structural interpolation, the back conversion from desired similarity constraints to a concrete structure is required. This technique enables generation of intermediate geometries for TS calculations (e.g., QST3) (Peng and Bernhard Schlegel, 1993; Peng et al., 1996), and it can also be used to guide BH searches of specified regions of the PES along isomerization pathways between two isomers. Furthermore, by implementing structural interpolation, one creates the opportunity to incorporate other GO techniques (e.g., particle swarm optimization) (Kennedy and Eberhart, 1995; Call et al., 2007; Shi et al., 2019) and machine learning techniques (e.g., growing neural gas) (Martinez and Schulten, 1991; Fritzke, 1994) into the BH algorithm. In practice, rather than an explicit analytical approach, structure interpolation can be achieved implicitly via local optimizations with a tolerable loss of accuracy. In our research, to efficiently use the nuclear configuration information from the BH simulation, we introduce both Euclidean distance matrix-based and cosine distance-based similarity functions together with the necessary techniques to accomplish structural interpolation. The mathematical and implementation details are described below.

Assessing Geometric Similarity

To begin assessing the similarity between two molecular geometries, one must first select an appropriate similarity function. One option, the Euclidean distance matrix representation ($\boldsymbol{D}$) of a molecule, is simply the collection of all interatomic distances as per (Gentle, 2007):

$$\mathbf{D}_{ij} = |\vec{r_i} - \vec{r_j}| \tag{2}$$

where $\vec{r_i}$ and $\vec{r_j}$ are the positional vectors (in Cartesian coordinates) of atoms i and j. Within the distance matrix representation, the similarity function is defined as the sum of the absolute difference between each atom pair for structures A and B:

$$d\left(\mathbf{D}_A, \mathbf{D}_B\right) = \sum i, j > i|\mathbf{D}_{A,ij} - \mathbf{D}_{B,ij}| \tag{3}$$

The distance matrix is a symmetric matrix with diagonal elements of zero. This representation is translationally and rotationally invariant, but not permutationally invariant (*viz.* identical nuclei are not necessarily chemically equivalent). Thus, in practice, the atom labeling should be adjusted such that the similarity index (the value of the similarity function) of the two input molecules is minimized. It should be noted that the memory requirement of this representation scales quadratically with the number of atoms. Consequently, the distance matrix approach is not a good choice for dealing with very large systems.

A second option is to represent the molecular nuclear configuration as a vector, $\vec{R}$ (Fu and Hopkins, 2018), containing the mass-weighted distance between each atom and the molecular center-of-mass:

$$\vec{R}_{COM} = \frac{\sum_i^{m_i} \vec{r_i}}{\sum_i m_i} \tag{4}$$

$$\vec{R_i} = m_i|\vec{r_i} - \vec{R}_{COM}| \tag{5}$$

Where m_i and $\vec{r}_i$ are the mass and the distance to the center-of-mass for the i^{th} atom. Given that the mass-weighted distance vector representation is in the center-of-mass frame, one can then calculate the cosine distance between the vectors for isomers A and B as per:

$$d(\vec{R_A}, \vec{R_B}) = \frac{\cos^{-1}\left(s(\vec{R_A}, \vec{R_B})\right)}{\pi} \tag{6}$$

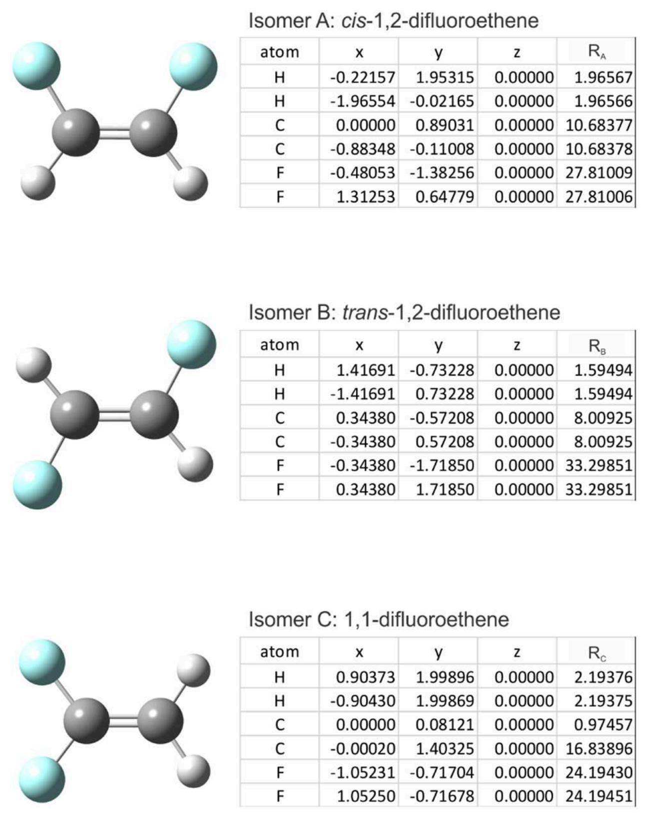
Where

$$s(\vec{R_A}, \vec{R_B}) = \frac{\vec{R_A} \cdot \vec{R_B}}{\left|\vec{R_A}\right|\left|\vec{R_B}\right|} \tag{7}$$

Again, this representation is translationally and rotationally invariant. However, care should be taken to ensure that the identity of the *i*th atom is retained throughout the BH search so that one compares the same atoms in each unique geometric structure. Alternatively, one might choose an operational convention whereby the resulting vector is sorted (e.g., smallest to largest values) prior to calculating cosine distance; this introduces a permutational invariance to the treatment for low symmetry systems. In contrast to the quadratic scaling of the distance matrix, the mass-weighted distance vector scales linearly with number of atoms. However, as a trade-off, the mass-weighted distance vector representation is less effective than the distance matrix approach in discriminating between conformers of highly symmetric species. For example, the mass-weighted distance vector representation is unable to distinguish square planar and tetrahedral conformations of methane given identical C–H bond length. Nevertheless, the uniqueness of the isomer-vector correspondence is still largely guaranteed in most cases in which only low symmetry structures are considered, particularly when relative energies are also considered in distinguishing isomeric/conformeric species.

The cosine similarity (Equation 7) ranges from −1 (meaning exactly opposite) to +1 (meaning identical). However, in practice, the cosine similarity for real molecular structures ranges from 0 to 1 since the center-of-mass vector is constructed from real space distances, which are always positive. Thus, two identical structures exhibit mass-weighted distance vectors with zero angular distance between them, and angular distances between vectors increase as the differences between the geometric structures of the associated isomers increase. For example, consider the isomers *cis*-1,2-difluoroethene, *trans*-1,2-difluoroethene, and 1,1-difluoroethene shown below in **Figure 2**. By inspection, one can identify that the mass-weighted distance vectors for the *cis*-1,2-difluoroethene and *trans*-1,2-difluoroethene isomers (R_A, R_B) are more like one another than they are to that of the 1,1-difluoroethene isomer (R_C). This is confirmed when calculating the cosine distances (see **Table 1**).

Calculating the distances between molecular structures facilitates analysis through agglomerative hierarchical clustering (Day and Edelsbrunner, 1984). This analysis provides a visual representation of the similarity of geometric structures—via production of a dendrogram plot—and therefore provides some insight into which species occupy similar regions of the potential energy landscape with respect to the mass-weighted nuclear coordinates. There are several methods available for analysis via agglomerative hierarchical clustering (Day and Edelsbrunner, 1984). One option for this analysis is the weighted pair group method with arithmetic mean (WPGMA), developed by Sokal and Michener (Michener and Sokal, 1957; Sokal and Michener, 1958). In each iteration of the WPGMA algorithm, the two nearest species (P and Q) are combined into a higher-level group P ∪ Q, thereby reducing the dimension of the $m \times m$ distance



Isomer A: *cis*-1,2-difluoroethene

atom	x	y	z	R_A
H	-0.22157	1.95315	0.00000	1.96567
H	-1.96554	-0.02165	0.00000	1.96566
C	0.00000	0.89031	0.00000	10.68377
C	-0.88348	-0.11008	0.00000	10.68378
F	-0.48053	-1.38256	0.00000	27.81009
F	1.31253	0.64779	0.00000	27.81006

Isomer B: *trans*-1,2-difluoroethene

atom	x	y	z	R_B
H	1.41691	-0.73228	0.00000	1.59494
H	-1.41691	0.73228	0.00000	1.59494
C	0.34380	-0.57208	0.00000	8.00925
C	-0.34380	0.57208	0.00000	8.00925
F	-0.34380	-1.71850	0.00000	33.29851
F	0.34380	1.71850	0.00000	33.29851

Isomer C: 1,1-difluoroethene

atom	x	y	z	R_C
H	0.90373	1.99896	0.00000	2.19376
H	-0.90430	1.99869	0.00000	2.19375
C	0.00000	0.08121	0.00000	0.97457
C	-0.00020	1.40325	0.00000	16.83896
F	-1.05231	-0.71704	0.00000	24.19430
F	1.05250	-0.71678	0.00000	24.19451

FIGURE 2 | The structures of **(top)** *cis*-1,2-difluoroethene, **(middle)** *trans*-1,2-difluoroethene, and **(bottom)** 1,1-difluoroethene. **(Inset Tables)** atomic coordinates and the mass-weighted distance vectors. Geometries were optimized at the PM6 level of theory as implemented in Gaussian 16 (Frisch et al., 2016).

TABLE 1 | The cosine distance matrix for *cis*-1,2-difluoroethene, *trans*-1,2-difluoroethene, and 1,1-difluoroethene.

Distance	*cis*-1,2-difluoro	*trans*-1,2-difluoro	1,1-difluoro
cis-1,2-difluoro	0	0.04200	0.09497
trans-1,2-difluoro	0.04200	0	0.10219
1,1-difluoro	0.09497	0.10219	0

Geometries were optimized at the PM6 level of theory as implemented in Gaussian 16 (Frisch et al., 2016).

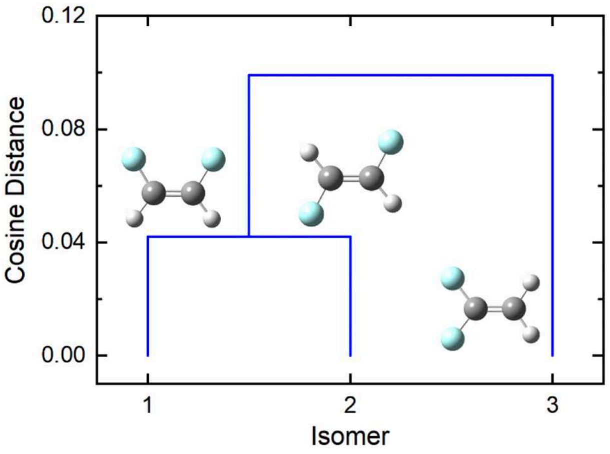


FIGURE 3 | The cosine distance dendrogram for difluoroethene. Molecular geometries were optimized at the PM6 level of theory as implemented in Gaussian 16 (Frisch et al., 2016).

matrix (e.g., **Table 1**) by one row and one column. The distance between group P ∪ Q and another group R is the arithmetic mean of the distances between the members of P ∪ Q and R, i.e.,:

$$d_{(P \cup Q),R} = \frac{d_{P,R} + d_{Q,R}}{2} \tag{8}$$

In the case of difluoroethene (**Figure 2** and **Table 1**), the smallest cosine distance of 0.042 between the *cis*- and *trans*-1,2-difluoroethene isomers would lead to their clustering as P ∪ Q, and the distance between this higher-level group and the 1,1-difluoroethene isomer would be (0.09497 + 0.10219)/2 = 0.09858. A dendrogram showing the hierarchical clustering of the isomers of difluoroethene is provided in **Figure 3**. By inspection of the dendrogram one can immediately see that the *cis*- and *trans*- isomers of 1,2-difluoroethene isomers are more closely related geometrically than either of these isomers is related to 1,1-difluoroethene.

Interpolating Intermediate Geometries

When searching complex PESs to find local minima or TSs, it is sometimes useful to interpolate geometries that are intermediate to two previously identified isomers. For example, consider the case in which a set of isomeric species has been identified, but one is very dissimilar from the others as determined by the geometric analysis described above. This might indicate that the BH search has become kinetically trapped and more attention should be paid to the region of the PES associated with the isolated structure. It is then useful to explore the PES between the more extensively mapped region and the region associated with the isolated structure to search for intermediates along the isomerization pathway and/or identify barriers to isomer interconversion. For the purpose of generating initial guess structures for the BH algorithm or for QST3 TS calculations, precise interpolation is not always necessary; (Peng and Bernhard Schlegel, 1993; Peng et al., 1996) most of the time interpolation can be accomplished implicitly, thereby improving the efficiency of the PES mapping. Currently, we have implemented two classes of implicit interpolation methods, one based on Monte Carlo sampling and the other based on molecular dynamics simulation.

Since the acceptance criteria are replaceable as a standard module in the evaluation part of the BH framework, instead of searching for low energy structures, one can choose to sample structures between two given minima on the PES within specified similarity constrains. Thus, a Monte Carlo with minimization approach can be established along a specified path/region of the PES. By applying an upper threshold to the distance of the sampled structure from the minima, one can constrain the search to a hyperdimensional ellipsoidal space between the two minima of interest. Within the distance matrix representation, the interpolation can also be accomplished with optimization on an interpolated artificial force field. Similar to the idea of the artificial force induced reaction (Maeda et al., 2014), the interpolated structure is obtained by minimizing a molecular mechanics-type force field, $\boldsymbol{V}$:

$$V(D_C) = \frac{\chi}{\bar{r}_{ij}} \left(D_{C,ij} - \bar{r}_{ij}\right)^2 \tag{9}$$

where χ is an arbitrary constant that facilitates optimization, and $D_{C,ij}$ and $\bar{r}_{ij}$, are the actual and expected interatomic distance of the interpolated structure. $\bar{r}_{ij}$ is constructed from the two minima, D_A and D_B and the interpolation factor, λ ($0 \leq \lambda \leq 1$) as per:

$$\bar{r}_{ij} = \lambda D_{A,ij} + (1 - \lambda) D_{B,ij} \tag{10}$$

The force field is thus a collection of harmonic terms whose force constant is inversely proportional to $\bar{r}_{ij}$. Compared to the Monte Carlo approach, using this force field approach in conjunction with standard geometry optimization techniques is expected to be more efficient at identifying intermediate structures owing to the reduced and more pertinent search space.

APPLICATION OF BH SEARCH RESULTS

Experimental measurements are typically concerned with probing ensembles, rather than single molecules. Consequently, it is necessary to identify which structures are present in the probed ensemble and the relative populations of those species. This can be particularly challenging for chemical systems that are kinetically trapped in a relatively high-energy region of the PES and for systems that are fluxional (i.e., those that can easily access multiple minima on the experimental time scale). To demonstrate the potential of our augmentation to the original BH method, we describe our efforts to model the infrared multiple photon dissociation (IRMPD) spectrum of proton-bound serine dimer and the temperature-depending collision cross section (CCS) of protonated alanine tripeptide, $[\text{AAA+H}]^+$.

Case Study 1: The IR Spectrum of the Protonated Serine Dimer

IRMPD spectroscopy has become one of the most effective techniques for determining the structure of molecular ions (Jašíková and Roithová, 2018). Ion spectra are recorded by isolating a specified *m/z* species in an ion trap and monitoring the fragmentation efficiency of the molecular ion as a function of the frequency of a probe laser, which passes through the ion trap, intersecting with the ion cloud (Lemaire et al., 2002; Oh et al., 2005; Polfer, 2011). Thus, IRMPD spectroscopy is a type of action spectroscopy whereby molecular fragmentation is interpreted as a signature of photon absorption. A detailed description of the technique is available in references (Aleese et al., 2006) and (Macaleese and Maître, 2007). By probing in the IR region, one obtains information on the frequencies of fundamental vibrational transitions, which may then be compared with the harmonic (and sometimes anharmonically-corrected) vibrational frequency predictions of electronic structure software packages. This, in turn, facilitates structural assignment based on the similarity between computed and measured spectra, and the identification of distinguishing/diagnostic spectral features.

Spectroscopic investigation of amino acids and amino acid-containing clusters continues to be an active field of research owing to the biological relevance of these systems (Nanita and Cooks, 2006; Mino et al., 2011; Stedwell et al., 2013; Sunahori et al., 2013; Armentrout et al., 2014; Seo et al., 2017, 2018; Heiles et al., 2018; Jašíková and Roithová, 2018; Ma et al., 2018; Scutelnic et al., 2018). In particular, serine has received a great deal of attention owing to the implication of the serine octamer in homochiral genesis (i.e., the origin of L-amino acid chiral preference in nature) (Counterman and Clemmer, 2001; Sunahori et al., 2013; Seo et al., 2017; Scutelnic et al., 2018). Indeed, the Bowers and von Helden groups recently published a series of high-profile studies detailing the assignment of the IR spectra for cryogenically-cooled protonated serine octamer, $[\text{Ser}_8 + \text{H}]^+$, and protonated serine dimer, $[\text{Ser}_2 + \text{H}]^+$ (Seo et al., 2017, 2018; Scutelnic et al., 2018). To demonstrate the utility of our augmented BH approach for searching PESs and assigning IR spectra, we employed our methodology to study $[\text{Ser}_2 + \text{H}]^+$.

To begin, preliminary B3LYP/6-311++G(d,p) optimizations were conducted for neutral and protonated serine monomers to obtain partial charges for utilization with the molecular mechanics force field. For neutral monomers, both canonical and zwitterionic initial guesses were employed, and only the canonical structures were obtained. For the protonated isomers, initial guesses protonated at the carbonyl group, the amine group, and the side-chain hydroxyl group were optimized; all resulted in an amine-protonated structure, in agreement with previously published results (Noguera et al., 2001). After the optimizations, the atomic partial charges were calculated using the CHelpG partition scheme to reproduce the electrostatic potential at the near exterior of the van der Waals radial surface (Breneman and Wiberg, 1990). DFT optimizations were run in parallel, threaded across 8 cores, and required approximately 1 hour per calculation. Following pre-optimization and partial charge calculations for the monomers, both moieties were combined to produce the protonated dimer for treatment with the BH code. To search the potential energy landscape, dihedral angles in both moieties were given random rotations of $-5^\circ \leq \phi \leq +5^\circ$ on each iteration of the BH algorithm. The neutral moiety was also given random rotations of $-5^\circ \leq \theta \leq +5^\circ$ around its body-fixed *x*–, *y*–, and *z* –axes, and random translation of -0.5 Å $\leq \eta \leq +0.5$ Å in each of the *x*–, *y*–, and *z*–directions. This ensures that the relative orientations of the two moieties are also sampled. For geometry optimization, the custom-written BH code interfaces with the Gaussian software package where the AMBER force-field is used as the model potential (Wang et al., 2006; Frisch et al., 2009). Following an initial run of 1,000 steps at a thermal energy of $E \approx 0.43$ eV ($T = 5{,}000$ K) to generate candidate structures, several parallel BH runs of 10,000 steps were run at a thermal energy of $E \approx 0.09$ eV ($T = 1{,}000$ K) to search the PES. In total, more than 60,000 cluster geometries were sampled.

To benchmark the augmented BH algorithm, eight standard BH simulations of 5,000 steps were conducted and structural

TABLE 2 | The results of eight (BH + interpolation) simulations of $[Ser_2 + H]^+$.

Simulation	# Isomers found		Global minimum (Hartree)	
	BH	+ Interpolation	BH	Interpolation
1	70	16	**−0.25984**	−0.25940
2	74	19	**−0.25984**	−0.25593
3	60	8	**−0.25984**	−0.25572
4	67	22	−0.25969	**−0.25984**
5	62	32	−0.25969	**−0.25984**
6	76	17	**−0.25984**	−0.25967
7	67	6	**−0.25984**	−0.25515
8	51	28	−0.25969	−0.25809

Geometries were optimized at the PM7 level of theory (Frisch et al., 2016). Bold values emphasize the lowest energy value among all the isomers obtained from the total of 8 searches.

interpolation was subsequently applied to the unique isomers identified at the PM7 level of theory. Unique $[Ser_2 + H]^+$ isomers were identified based on energetic differences ($\Delta E \geq 10^{-5}$ Hartree) and by using a value of 50.0 Å as the similarity threshold between isomer pairs within the Euclidean distance matrix (*vide supra*). Isomer pairs with Euclidean distances of more than 150.0 Å were candidates for structural interpolation. Due to the large number of potential isomer pairs (~6,000 for each BH simulation), we chose to randomly select only 300 pairs to test the interpolation methodology. For each pair, the midpoint structure ($\lambda = 0.5$) was located as described above and optimized at the PM7 level of theory. The optimized geometry of the interpolated structure was then compared to those in the original BH set using the same energy and Euclidean distance thresholds as employed previously. The results of the eight parallel (BH + interpolation) simulations are summarized in **Table 2**.

There are two observations worth noting in **Table 2**. Firstly, the isomer sets that were identified by the standard BH algorithm are augmented considerably by post-simulation interpolation; on average 19 new isomers were identified by interpolating between the 300 randomly selected isomer pairs found by standard BH simulations. Secondly, although the global minimum structure was identified in only five of the eight standard BH simulations of 5,000 steps, introducing post-simulation interpolation improved the rate of identifying the $[Ser_2 + H]^+$ global minimum to seven out of eight simulations.

Following BH simulation, the 200 unique lowest energy structures were carried forward to re-optimization at the B3LYP/6-311++G(d,p) + GD3 level of theory (Becke, 1988, 1993; Grimme et al., 2010). This treatment reduced the total number of unique isomers to 40. To ensure that these structures were local minima on the PES (i.e., no negative eigenvalues in the Hessian matrix, rather than TSs which have one negative Hessian eigenvalue), harmonic frequency calculations were undertaken. These calculations also served to predict the vibrational (*viz.* IR) spectra of the isomers and to estimate thermochemical corrections (see sections 1.1 and 1.2 of the **Supplementary Materials** for details). Using the optimized geometries from the density functional theory calculations, the distance matrix (as described in Equations 2, 3) was constructed. Linkages for hierarchical clustering were then determined using Ward's minimum variance method as implemented in the Orange software package (https://orange.biolab.si/) (Demsar et al., 2013), which at each step finds the pair of clusters that leads to the minimum increase in total within-cluster variance after merging (Ward, 1963). The resulting dendrogram, which is plotted in **Figure 4**, clearly shows four distinct groups of geometric structures; these groups are highlighted in blue, red, green, and orange. To better visualize the data, we have also used multi-dimensional scaling to create a 2D plot of the clustered data (Wickelmaier, 2003; Borg and Groenen, 2005). Based on this hierarchical clustering analysis, we clearly see that the BH algorithm identified several local minima associated with four distinct regions of the $[Ser_2 + H]^+$ PES. The lowest energy isomer in each of these four regions (*viz.* isomers 1, 6, 14, and 22) are highlighted and labeled on the MDS plot. This type of analysis provides insight with respect to how thoroughly a region of the PES has been searched. For example, if only one or two data points were identified in the blue region of the MDS plot, one might decide to initialize an additional BH run starting from one of the previously identified geometries. Moreover, this analysis can help guide interpolation efforts to identify TSs or geometries associated with stable intermediates between two previously identified minima. For example, upon inspection of the MDS plot shown in **Figure 4**, one can identify two outliers associated with the red group (in the top left of the red section) and one outlier associated with the green group (bottom left of the green section). In principle, one might choose to explore the region between these features and the more closely clustered structures on the MDS plot via the methods described in section Interpolating Intermediate Geometries. We choose not to do so here, however, because these three structures are associated with isomers 38, 39, and 40 (the highest energy species in our set).

Having identified four low energy geometric groupings associated with the $[Ser_2 + H]^+$ PES, we can then visually inspect the structures to rationalize their association via hierarchical clustering. In doing so, we find that the clustered species are associated with four distinct binding motifs, which we label motifs 1 (orange), 2 (blue), 3 (green), and 4 (red). The 3D structures and 2D chemical structures for the lowest energy isomer in each group is provided in **Figure 5**. Motifs 1 and 3 are associated with bidentate complexation between the ammonium group of the protonated moiety and the neutral moiety. In the case of motif 1, the ammonium group forms intermolecular hydrogen bonds with the amino group and the hydroxyl group of the neutral moiety. In contrast, motif 3 forms intermolecular hydrogen bonds with the hydroxyl group and the carboxylic acid group of the neutral moiety. Motifs 2 and 4 are associated with monodentate complexation between the ammonium group of the protonated moiety and the neutral moiety. These two binding motifs differ in terms of the relative orientations of the two serine moieties and with respect to the presence of a O–H•••N intramolecular hydrogen bond (IMHB) in the neutral moiety (motif 2) versus a O–H•••O IMHB in the neutral moiety (motif 4).

To determine which (if any) of the computed $[Ser_2 + H]^+$ isomers are observed experimentally, calculated harmonic vibrational spectra were compared against the experimental

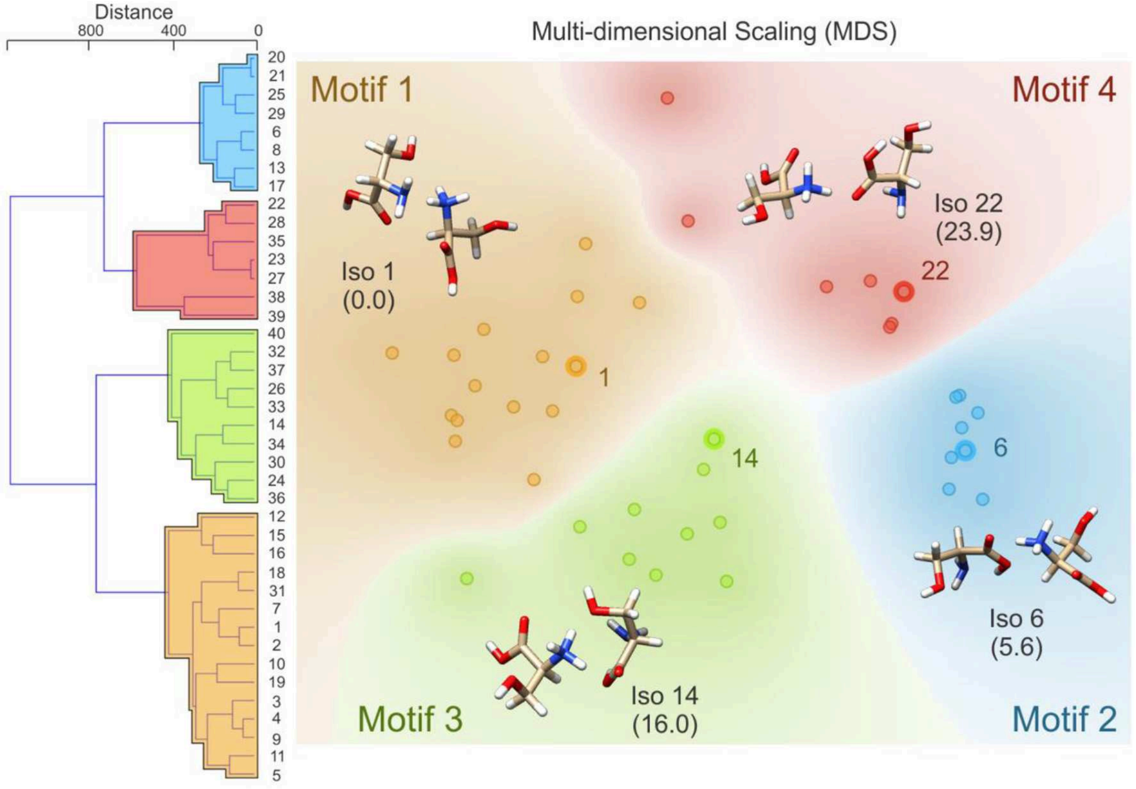


FIGURE 4 | (Left) The distance dendrogram for the protonated serine dimer. Isomer numbers are indicated for each branch of the dendrogram. **(Right)** A multi-dimensional scaling 2D projection of the hierarchical clustered data. Isomers are numbered in order of increasing energy above the global minimum (isomer 1). Standard Gibbs energies (in parentheses) are reported in kJ mol^{-1}. Calculations were conducted at the B3LYP/6-311++G(d,p) + GD3 level of theory as implemented in Gaussian 09 (Frisch et al., 2009).

IRMPD spectrum using the methodology outlined by Fu and Hopkins (2018) The experimental spectrum employed was a concatenation of the spectra recorded by Seo et al. in the 1,000–1,900 cm^{-1} region and by Sunahori et al. in the 3,200–3,800 cm^{-1} region (Sunahori et al., 2013; Seo et al., 2018). These spectra were digitized using a custom-written python script from figures in their respective publications, interpolated in 2 cm^{-1} intervals, then normalized such that the maximum intensity in each region was set to 1. Calculated IR spectra were first scaled using appropriate frequency scaling factors and broadened with a Lorentzian line shape of 15 cm^{-1} FWHM (Andersson and Uvdal, 2005; Fu and Hopkins, 2018), and then were similarly interpolated and normalized. The intensity vectors (i.e., y-values) of the computed spectra were then compared with the experimental spectrum by taking the Euclidian distance (d_{Euc}) between the intensity vectors and assigning a scaled similarity index as per:

$$Scaled\ Similarity = 1 - \frac{\left(d_{Euc} - d_{Euc}^{Min}\right)}{\left(d_{Euc}^{Max} - d_{Euc}^{Min}\right)} \qquad (11)$$

Where d_{Euc}^{Min} is the minimum Euclidean distance amongst the set of vectors and d_{Euc}^{Max} is the maximum Euclidean distance amongst the set of vectors following subtraction of the minimum distance. This treatment generates a scaled similarity index that ranges between 0 (worst match) and 1 (best match). The scaled similarities for the computed $[Ser_2 + H]^+$ isomer spectra are plotted in **Figure 6**. Inspection of **Figure 6** indicates that Isomer 6 yields a significantly better match to the experimental spectrum than do other isomers. Moreover, we find that four of the five best matches are provided by isomers associated with binding motif 2. This suggests that, despite the fact that motif 1 is associated with the lowest energy region of the $[Ser_2 + H]^+$ PES at $T = 298$ K and $P = 1$ atm, the region of the PES associated with motif 2 is predominantly populated in ion trap experiments.

Figure 7 plots the experimental IRMPD spectrum for $[Ser_2 + H]^+$ and the computed spectra for isomers 1, 6 (best match), 14, and 22—the lowest energy isomers associated with each of the four binding motifs. The diagnostic peaks, which are highlighted in blue in **Figure 7**, are associated with the HNH angle bending motions (*ca.* 1,450 cm^{-1}) and N–H bond stretching motions (*ca.* 3,250 cm^{-1}) of the ammonium and amino groups. Although isomer 1 is the global minimum structure based on standard Gibbs energies, the spectrum of isomer 6 (+5.6 kJ mol^{-1}) is much more representative of the experimental spectrum. This was also noted by Sunahori et al., who identified isomer 6 in their study (Sunahori et al., 2013). Kong et al. also identified isomer 6 in

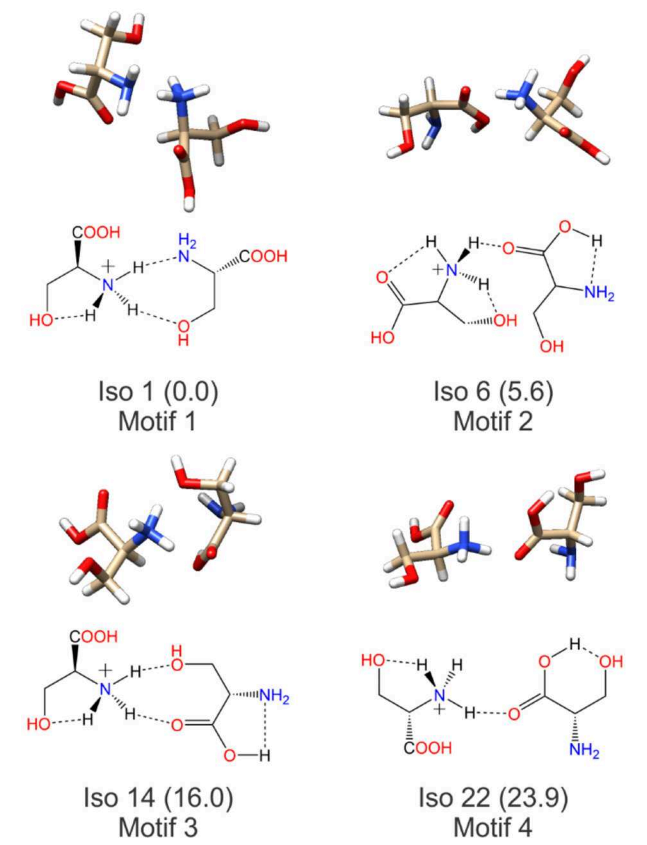


FIGURE 5 | The lowest energy isomers for each low energy binding motif of the protonated serine dimer. Motifs 1 and 3 show bidentate coordination between the two moieties, whereas motifs 2 and 4 exhibit monodentate coordination between the two moieties.

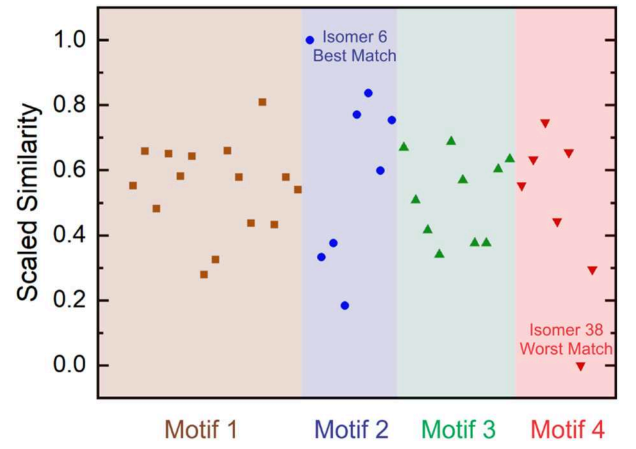


FIGURE 6 | Scaled Euclidean similarities of computed harmonic vibrational spectra to experimental IRMPD spectra for the protonated serine dimer. Isomer 6 gives the best match and Isomer 38 gives the worst match amongst the 40-isomer set. Isomers are ordered in increasing energy from left to right in each motif.

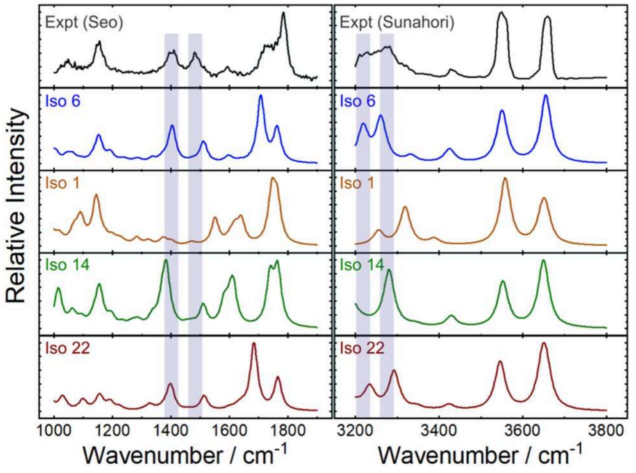


FIGURE 7 | Experimental IRMPD spectra and computed harmonic vibrational spectra for the protonated serine dimer. The experimental spectra were adapted from Seo et al. (2018) and Sunahori et al. (2013). The computed IR spectra are associated with the lowest energy isomer for each of the four binding motifs. Scaling factors of 0.9679 and 0.95 were employed for the 1,000–1,900 cm^{-1} and 3,200–3,800 cm^{-1} regions, respectively (Andersson and Uvdal, 2005; Fu and Hopkins, 2018).

their work, but apparently did not consider it in their spectral assignment (Kong et al., 2006). Note that harmonic spectra were scaled by 0.9679 in the 1,000–2,000 cm^{-1} region and 0.95 in the 3,000–4,000 cm^{-1} region, as recommended by NIST and based on previous work for similar systems (Andersson and Uvdal, 2005; Fu and Hopkins, 2018).

It is necessary to highlight three caveats for the above example of identifying the spectral carrier of $[\mathrm{Ser}_2 + \mathrm{H}]^+$. First, to create the experimental spectrum that we used in our assignment, we collated the results of two separate studies (Sunahori et al., 2013; Seo et al., 2018). It is not necessarily true that the same ensemble populations were produced under the experimental conditions employed in both of these studies. However, given that isomer 6 provides the best match to both regions of the experimental spectrum, it seems to be that instrument conditions were similar in these two cases. A second consideration is the fact that peak intensities in IRMPD spectra are not necessarily well-modeled by computed absorption spectra owing to the fact that IRMPD intensities are dependent on absorption cross sections *and* the coupling efficiency for accessing dissociative channels. (Parneix et al., 2013) The methodology outline above assumes that the computed linear absorption intensities are representative of IRMPD intensities or, barring that, that the IRMPD intensities for a given band vary similarly from the computed intensity for all isomeric species. Finally, the above treatment also assumes that the computed harmonic frequencies suitably model the experimental spectrum. The validity of this assumption depends on the accuracy of the model chemistry and on the anharmonicity of the system being studied. While the $[\mathrm{Ser}_2 + \mathrm{H}]^+$ is apparently well-modeled by the B3LYP/6-311++G(d,p) + GD3 approach employed here, one should in general be aware of the anharmonic nature of hydrogen bonds and shared protons (Schofield et al., 2005; Oomens et al., 2009; Steill et al., 2011; Ieritano et al., 2016).

Case Study 2: Dynamic Collision Cross Section of Protonated Alanine Tripeptide

Ion mobility spectrometry (IMS) is widely employed in the detection of illicit substances and for structural elucidation of ions (Collins and Lee, 2002; Verkouteren and Staymates, 2011; Lapthorn et al., 2013; Lanucara et al., 2014; Cumeras et al., 2015; Cajka and Fiehn, 2016; Paglia and Astarita, 2017). The success of IMS in determining analyte structure relies on accurate modeling of ion structure and subsequent calculation of CCSs for comparison with those determined experimentally. Experimental CCSs are obtained by relating the ion mobility, K, to CCS via the Mason-Schamp Equation (Mason and Mcdaniel, 1988; Ieritano et al., 2019b):

$$K = \frac{\sqrt{18\pi}}{16}\sqrt{\frac{1}{m_{ion}} + \frac{1}{m_{gas}}}\frac{ze}{\sqrt{k_b T}}\frac{1}{\Omega_{avg}}\frac{1}{N} \tag{12}$$

Where m_{gas} is the mass of the buffer gas, N is the number density of the gas, m_{ion} is the mass of the ion, z is the ion charge state, e is the elementary charge, k_b is the Boltzmann constant, T is the temperature, and Ω_{avg} is the orientationally-averaged CCS. Typically, ion structures are viewed as rigid and ensembles are approximated as being composed of only a single structure in cases where multiple distinct signals are unresolved. This view is somewhat tenuous, particularly in the differential mobility spectrometry (DMS) variant of IMS wherein rapidly oscillating electric field conditions drive separations based on mobility differences between the high- and low-field portions of the applied waveform (Guevremont and Purves, 1999; Guevremont, 2004; Krylov et al., 2007, 2009; Krylov and Nazarov, 2009; Hopkins, 2015, 2019). The phenomenon of differential ion mobility is still not well-understood, and there is as yet no first principles model (Guevremont and Purves, 1999; Guevremont, 2004; Krylov et al., 2007, 2009; Krylov and Nazarov, 2009; Hopkins, 2015, 2019). However, one can view the effective temperature of an analyte ion in terms of the changing field conditions; the ion is relatively cold under low-field conditions and relatively hot under high-field conditions (Viehland and Mason, 1995; Robinson et al., 2008; Hopkins, 2019). By estimating ion temperatures with two-temperature theory (Robinson et al., 2008; Siems et al., 2016), we find that field-induced heating leads to effective ion temperature variations in the range of 300–800 K during one duty cycle of the commonly applied maximum electric field in the DMS cell (Hopkins, 2019). The variation in electric field, and therefore effective ion temperature, affects the ion mobility in two ways, the most obvious being the reduction of mobility with increasing temperature as predicted by Equation (12). Somewhat more subtle is the fact that Ω_{avg} must also be temperature-dependent since at elevated temperatures ions are able to access a larger region of the associated PES (assuming equipartition amongst the various DoFs of the molecule). Consequently, to accurately model an ion's Ω_{avg}, one must identify which geometric structures are accessible under the given experimental conditions and estimate the contribution of that structure to the time-averaged CCS of the ion.

If we consider the case of protonated alanine tripeptide, $[\mathrm{AAA} + \mathrm{H}]^+$, there are several internal DoFs associated with dihedral angle rotations that can yield a variety of conformations. Upon application of the BH algorithm to search the PES of the $[\mathrm{AAA} + \mathrm{H}]^+$ molecular ion, followed by re-optimization of the candidate structures at the B3LYP/6-31++G(d,p) + GD3 level of theory (Becke, 1988, 1993; Grimme et al., 2010), fourteen low energy conformations were identified. These structures are shown in **Figure 8** along with their relative standard Gibbs energies (in kJ mol^{-1}) (see sections 2.1 and 2.2 of the **Supplementary Materials** for details). Calculating the cosine distances between the various mass-weighted distance vectors and subsequent application of WPGMA agglomerative hierarchical clustering yields the dendrogram plot shown in **Figure 8**. Five unique sets of conformers are highlighted in the dendrogram. The set highlighted in yellow, of which the global minimum conformer is a member, contains compact structures that are stabilized by an IMHB between the protonated N-terminus and the carbonyl oxygen atom of the C-terminus. The set highlighted in green also contains relatively compact structures, but hydrogen bonding instead occurs between the protonated N-terminus and the hydroxyl group of the C-terminus. The set highlighted in red, on the other hand, contains elongated structures (i.e., the N- and C-termini do not interact). Conformers 6 and 9 (blue and orange, respectively) are intermediate species between the compact species (yellow and green sets) and the elongated species (red set). In the case of conformer 6, the N-terminus forms an IMHB with the nearest amide carbonyl rather than with the C-terminus. In contrast, the C-terminus of conformer 9 forms an IMHB with the most distant amino nitrogen instead of with the N-terminus.

If we calculate the relative Gibbs energies of the $[\mathrm{AAA} + \mathrm{H}]^+$ conformers as a function of temperature, an interesting picture emerges. Owing to differences in the entropic contributions to the Gibbs energies, at low temperature the compact, H-bonded conformers associated with the yellow group are the dominant species in the ensemble, whereas at high temperature the elongated, non-H-bonded species in the red group dominate. One can estimate the relative populations of the various conformers via (Oh and Zeng, 1999; Vehkamäki, 2006; Hopkins, 2019):

$$N_i = N_0 e^{-\frac{\Delta G_{rel}}{k_B T}} \tag{13}$$

Where N_0 is the relative population of the lowest energy cluster (usually set to 1), N_i is the relative population of the ith cluster, ΔG_{rel} is the Gibbs energy of formation relative to the lowest energy cluster, and k_B is Boltzmann's constant. By calculating the relative populations of the clusters as a function of temperature (at a constant pressure of P = 1 atm), one can produce a temperature-dependent relative population plot as shown in **Figure 9**.

Figure 9 shows that at *ca*. T = 420 K $[\mathrm{AAA} + \mathrm{H}]^+$ conformer 3 becomes the most populated species in the ensemble (i.e., the global minimum structure on the Gibbs energy surface). As

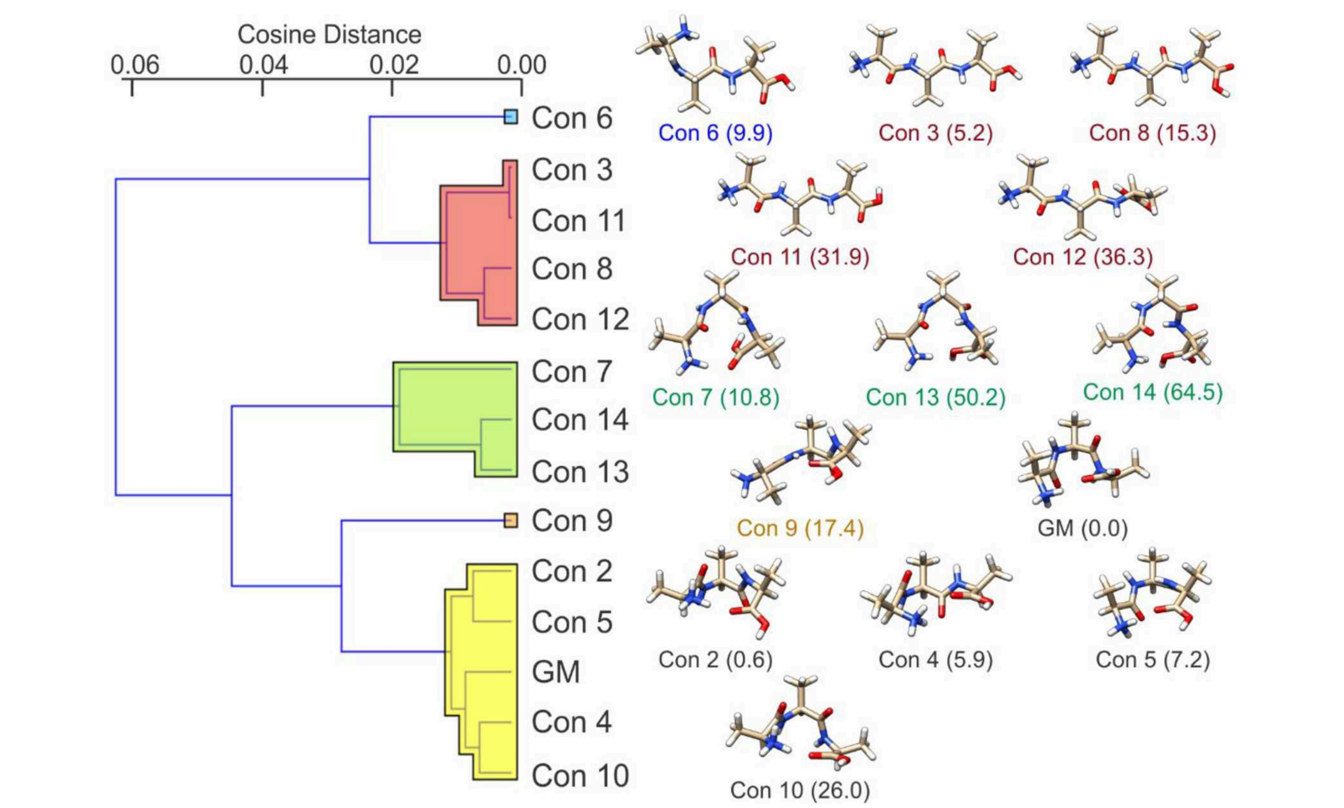


FIGURE 8 | (Left) The cosine distance dendrogram for protonated alanine tripeptide, $[AAA+H]^+$. **(Right)** Molecular geometries and relative standard Gibbs energies (kJ mol^{-1}; in parentheses). Calculations were conducted at the B3LYP/6-311++G(d,p) + GD3 level of theory as implemented in Gaussian 16 (Frisch et al., 2016). Conformers are numbered in order of increasing energy relative to that of the global minimum (GM) structure.

the temperature increases further, conformer 3 is increasingly stabilized with respect to conformers 1 (the low T global minimum) and 2. At temperatures above 660 K, conformers 1 and 2 become minor contributors to the overall ensemble population in favor of conformers 3 and 8. This "tilting" of the Gibbs energy landscape as a function of temperature essentially decants the conformers associated with the yellow set into the red set (see **Figure 8**) as field-induced ion temperature increases, and back again as the temperature decreases during the low field portion of the oscillating DMS waveform. This dynamic process of peptide unfolding and re-folding yields a dynamic temperature-dependent ion CCS that, along with the effect of increased carrier gas viscosity at higher temperature (Mason and Mcdaniel, 1988; Hopkins, 2019), gives rise to differential mobility behavior. If one assumes that the ion quickly reaches thermal equilibrium, which is likely given the conditions of the DMS cell (1 atm of carrier gas), one can estimate the temperature-dependent ion CCS as a sum of the Boltzmann-weighted conformer CCSs (Ieritano et al., 2019a). This is plotted for $[AAA + H]^+$ in **Figure 10**. It is worth noting that the experimentally-measured T ≈ 293 K value of $\Omega_{ave}(N_2) = 151$ Å^2 (Bush et al., 2012) is well-modeled by the T = 300 K Boltzmann-weighted sum of the various isomer CCSs as calculated using the MobCal-MPI code (https://uwaterloo.ca/hopkins-lab/mobcal-mpi), $\Omega_{Boltzmann}(N_2) = 151.3$ Å^2 (Ieritano et al., 2019b). In comparison, the calculated CCS for the static global minimum structure is $\Omega_{Boltzmann}(N_2) = 148.7$ Å^2. This demonstrates that even at a relatively low fixed temperature, there is some benefit in considering the relative populations of the conformeric species present in the experimental ensemble.

SUMMARY

Because the PESs of complex, fluxional molecular systems tend to be characterized by multiple funnels (*viz.* collections of closely related local minima), the BH framework has proven to be an effective search and optimization strategy (Locatelli, 2005; Olson et al., 2012). However, owing to the stochasticity of the algorithm, which is predominantly due to the random perturbative component, it is sometimes useful to introduce additional criteria which limit the regions of exploration on the PES. This has been traditionally accomplished by exploring specific degrees of freedom (e.g., dihedral rotations) on the potential energy landscape and by introducing a thermal energy distribution as a probabilistic means of accepting/rejecting random geometric perturbations. We have also introduced techniques from unsupervised machine learning, specifically distance matrices and hierarchical clustering, to further augment the BH algorithm. Although currently implemented as a separate module, these machine learning augmentations will in the

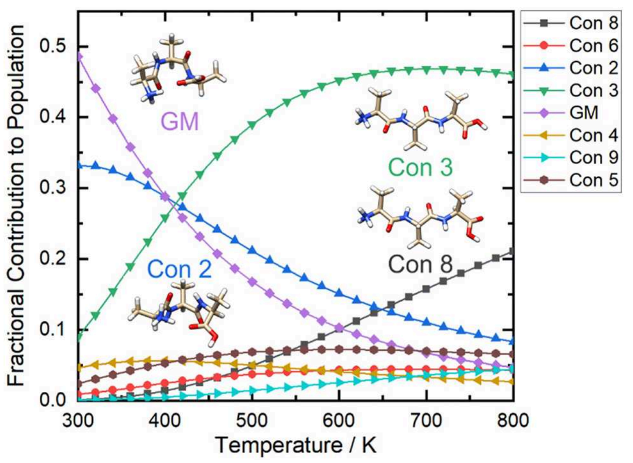


FIGURE 9 | The relative populations of the low energy conformers of protonated alanine tripeptide, $[AAA{+}H]^+$, as estimated via Gibbs energy calculations over the temperature range T = 300–800 K. Calculations were conducted at the B3LYP/6-311++G(d,p) + GD3 level of theory as implemented in Gaussian 16 (Frisch et al., 2016). Conformers are numbered in order of increasing energy relative to that of the global minimum (GM; i.e., conformer 1) structure.

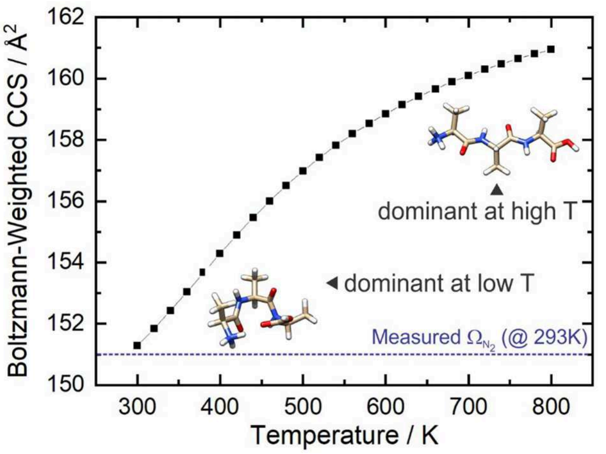


FIGURE 10 | The Boltzmann-weighted CCS of $[AAA + H]^+$ as a function of temperature at P = 1 atm. The dashed blue line shows the orientationally-averaged CCS, Ω_{ave}, measured in N_2 at room temperature (T ≈ 293 K) (Bush et al., 2012).

future be incorporated for on-the-fly geometric analyses, which would ultimately provide additional control and efficiency during execution of the search algorithm afforded by reducing the search space to pertinent regions connecting known stationary points. This is particularly useful in identifying intermediate local minima and TSs between known isomers. Moreover, utilizing these same methods post-BH provides deep insights into the relation between stationary points and how these are partitioned on the potential energy landscape. This can be of great benefit in modeling experimental ensembles and in rationalizing the observation of kinetically-trapped species and dynamic molecular geometries.

In this manuscript we highlight the power of the BH framework in two case studies: (1) assigning the spectral carrier(s) of the IRMPD spectrum of $[Ser_2{+}H]^+$ and (2) modeling the temperature-dependent collision cross sections of $[AAA{+}H]^+$. In case study 1, we show that a thorough mapping of the potential energy landscape is warranted to identify the species probed in gas phase ion spectroscopic studies of weakly-bound clusters. In the case of the protonated serine dimer, rather than observing the lowest energy isomer (as expected based on standard Gibbs energies), Seo et al. and Sunahori et al. observed a species that was associated with a relatively remote, higher energy region of the cluster PES (Sunahori et al., 2013; Seo et al., 2018). It is still an open question as to whether this was due to kinetic trapping during production or formation of this species *in situ* due to field-induced heating within the ion traps. In case study 2, we show that mapping PESs to identify low energy conformer geometries, which were subsequently refined at a higher level of quantum chemical theory, provides insight into how molecular geometry changes with increasing temperature. For $[AAA{+}H]^+$, increasing the temperature of the system results in the dissociation of IMHBs and the formation of larger elongated structures compared to the compact H-bonded species favored at low temperature. We also demonstrate that modeling molecular collision cross sections as a Boltzmann-weighted sum of the CCSs for accessible conformers provides an accurate estimate of those measured experimentally (0.3 Å^2 difference). It should be noted that this treatment assumes that the accessible conformers are readily interconvertible, and that thermal equilibrium is quickly established. In principle, one could also employ the interpolation techniques described in section Interpolating Intermediate Geometries to calculate barriers to interconversion and validate this assumption. However, the fact that our calculations yield results that are in excellent agreement with experimental measurements indicates that, in this case, the assumption is valid.

Ultimately, the BH framework is a useful approach to characterizing the structures and dynamics of chemical systems which exhibit PESs of high dimensionality. Examples of such systems range from weakly-bound nanoclusters to biological macromolecules. We note that, despite the success of our current implementation, the development of the BH framework by ourselves and others is ongoing. We expect that further tuning will improve general performance and, owing to the versatility of the method, that BH performance for specific tasks will continue to improve by tailoring key features of the algorithm.

AUTHOR CONTRIBUTIONS

CZ conducted the serine dimer work and wrote the original draft of manuscript. CI conducted the alanine tripeptide work. WH wrote the final draft of the manuscript.

FUNDING

WH acknowledges funding from the Natural Sciences and Engineering Research Council (NSERC) of Canada in the form of a Discovery Grant and from the Province of Ontario in the form of an Early Researcher Award (ERA). CI acknowledges funding from NSERC in the form of a post graduate scholarship.

ACKNOWLEDGMENTS

The authors would like to acknowledge Amy M. -C. Yang for help with data acquisition.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2019.00519/full#supplementary-material

Calculated energies and atomic XYZ coordinates for all species are available as **Supplementary Material**.

REFERENCES

Aleese, L. M., Simon, A., Mcmahon, T. B., Ortega, J.-M., Scuderi, D., Lemaire, J., et al. (2006). Mid-IR spectroscopy of protonated leucine methyl ester performed with an FTICR or a Paul type ion-trap. *Int. J. Mass Spectrometry* 249–250, 14–20. doi: 10.1016/j.ijms.2006.01.008

Andersson, M. P., and Uvdal, P. (2005). New scale factors for harmonic vibrational frequencies using the B3LYP density functional method with the triple-ζ basis set 6-311+G(d,p). *J. Phys. Chem. A* 109, 2937–2941. doi: 10.1021/jp045733a

Armentrout, P. B., Yang, B., and Rodgers, M. T. (2014). Metal cation dependence of interactions with amino acids: Bond dissociation energies of Rb+ and Cs+ to the acidic amino acids and their amide derivatives. *J. Phys. Chem. B* 118, 4300–4314. doi: 10.1021/jp5001754

Becke, A. D. (1988). Density-functional exchange-energy approximation with correct asymptotic-behavior. *Phys. Rev. A* 38, 3098–3100. doi: 10.1103/PhysRevA.38.3098

Becke, A. D. (1993). Density-functional thermochemistry.3. the role of exact exchange. *J. Chem. Phys.* 98, 5648–5652. doi: 10.1063/1.464913

Borg, I., and Groenen, P. J. F. (2005). *Modern Multidimensional Scaling.* New York, NY: Springer.

Breneman, C. M., and Wiberg, K. B. (1990). Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. *J. Comput. Chem.* 11, 361–373. doi: 10.1002/jcc.540110311

Bush, M. F., Campuzano, I. D. G., and Robinson, C. V. (2012). Ion mobility mass spectrometry of peptide ions: effects of drift gas and calibration strategies. *Analyt. Chem.* 84, 7124–7130. doi: 10.1021/ac3014498

Cajka, T., and Fiehn, O. (2016). Toward merging untargeted and targeted methods in mass spectrometry-based metabolomics and lipidomics. *Analyt. Chem.* 88, 524–545. doi: 10.1021/acs.analchem.5b04491

Call, S. T., Zubarev, D. Y., and Boldyrev, A. I. (2007). Global minimum structure searches via particle swarm optimization. *J. Comput. Chem.* 28, 1177–1186. doi: 10.1002/jcc.20621

Campbell, J. L., Le Blanc, J. C. Y., and Schneider, B. B. (2012). Probing electrospray ionization dynamics using differential mobility spectrometry: the curious case of 4-aminobenzoic acid. *Analyt. Chem.* 84, 7857–7864. doi: 10.1021/ac301529w

Campbell, J. L., Yang, A. M.-C., Melo, L. R., and Hopkins, W. S. (2016). Studying gas-phase interconversion of tautomers using differential mobility spectrometry. *J. Am. Soc. Mass Spectrom.* 27, 1277–1284. doi: 10.1007/s13361-016-1392-2

Cereto, M. (2015). Molecular fingerprint similarity search in virtual screening. *Methods* 71, 58–63. doi: 10.1016/j.ymeth.2014.08.005

Choudhary, B. (2003). *The Elements of Complex Analysis, 2nd Edn*, ed K. K. Gupta (New Delhi: New Age International Limited Publishers).

Collins, D., and Lee, M. (2002). Developments in ion mobility spectrometry–mass spectrometry. *Analyt. Bioanalyt. Chem.* 372, 66–73. doi: 10.1007/s00216-001-1195-5

Counterman, A. E., and Clemmer, D. E. (2001). Magic number clusters of serine in the gas phase. *J. Phys. Chem. B* 105, 8092–8096. doi: 10.1021/jp011421l

Cumeras, R., Figueras, E., Davis, C. E., Baumbach, J. I., and Gràcia, I. (2015). Review on ion mobility spectrometry. Part 1: current instrumentation. *Analyst* 140, 1376–1390. doi: 10.1039/C4AN01100G

Day, W. H. E., and Edelsbrunner, H. (1984). Efficient algorithms for agglomerative hierarchical clustering methods. *J. Classification* 1, 7–24. doi: 10.1007/BF01890115

Demsar, J., Curk, T., Erjavex, A., Gorup, C., Hocevar, T., Milutinovic, M., et al. (2013). Orange: data mining toolbox in python. *J. Mach. Learn. Res.* 14, 2349–2353.

Eberhart, R., and Yuhui, S. (2001). "Particle swarm optimization: developments, applications and resources," in *Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No.01TH8546),* 81, 81–86.

Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., et al. (2009). *Gaussian 09 Revision D.01.* Wallingford, CT: Gaussian, Inc.

Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., et al. (2016). *Gaussian 16 Rev. B.01.* Wallingford, CT.

Fritzke, B. (1994). "NIPS'94," in *Proceedings of the 7th International Conference on Neural Information Processing Systems*. Cambridge, MA: MIT Press, 625–632.

Fu, W., and Hopkins, W. S. (2018). Applying machine learning to vibrational spectroscopy. *J. Phys. Chem. A* 122, 167–171. doi: 10.1021/acs.jpca.7b10303

Gentle, J. E. (2007). *Matrix Algebra Theory, Computations, and Applications in Statistics*. New York, NY: Springer, 261–319.

Grimme, S., Antony, J., Ehrlich, S., and Krieg, H. (2010). A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* 132:154104. doi: 10.1063/1.3382344

Guevremont, R. (2004). High-field asymmetric waveform ion mobility spectrometry: a new tool for mass spectrometry. *J. Chromatogr. A* 1058, 3–19. doi: 10.1016/S0021-9673(04)01478-5

Guevremont, R., and Purves, R. W. (1999). Atmospheric pressure ion focusing in a high-field asymmetric waveform ion mobility spectrometer. *Rev. Sci. Instruments* 70, 1370–1383. doi: 10.1063/1.1149599

Heiles, S., Berden, G., Oomens, J., and Williams, E. R. (2018). Competition between salt bridge and non-zwitterionic structures in deprotonated amino acid dimers. *Phys. Chem. Chem. Phys.* 20, 15641–15652. doi: 10.1039/C8CP01458B

Heller, S., Mcnaught, A., Stein, S., Tchekhovskoi, D., and Pletnev, I. (2013). InChI - The worldwide chemical structure identifier standard. *J. Cheminform.* 5, 1–9. doi: 10.1186/1758-2946-5-7

Hopkins, W. S. (2015). Determining the properties of gas-phase clusters. *Mol. Phys.* 113, 3151–3158. doi: 10.1080/00268976.2015.1053545

Hopkins, W. S. (2019). "Chapter four - dynamic clustering and ion microsolvation," in *Comprehensive Analytical Chemistry, Vol. 83*, eds. W.A. Donald and J.S. Prell (Amsterdam: Elsevier), 83–122.

Hopkins, W. S., Marta, R. A., and Mcmahon, T. B. (2013). Proton-bound 3-cyanophenylalanine trimethylamine clusters: isomer-specific fragmentation pathways and evidence of gas-phase zwitterions. *J. Phys. Chem. A* 117, 10714–10718. doi: 10.1021/jp407766j

Hopkins, W. S., Marta, R. A., Steinmetz, V., and Mcmahon, T. B. (2015). Mode-specific fragmentation of amino acid-containing clusters. *Phys. Chem. Chem. Phys.* 17, 28548–28555. doi: 10.1039/C5CP03517A

Ieritano, C., Campbell, J. L., and Hopkins, W. S. (2019a). Unravelling the factors that drive separation in differential mobility spectrometry: a case study of regioisomeric phosphatidylcholine adduct. *Int. J. Mass Spectrom.* 444:116182. doi: 10.1016/j.ijms.2019.116182

Ieritano, C., Carr, P. J. J., Hasan, M., Burt, M., Marta, R. A., Steinmetz, V., et al. (2016). The structures and properties of proton- and alkali-bound cysteine dimers. *Phys. Chem. Chem. Phys.* 18, 4704–4710. doi: 10.1039/C5CP07414B

Ieritano, C., Crouse, J., Campbell, J. L., and Hopkins, W. S. (2019b). A parallelized molecular collision cross section package with optimized accuracy and efficiency. *Analyst* 144, 1660–1670. doi: 10.1039/C8AN02150C

Jašíková, L., and Roithová, J. (2018). Infrared multiphoton dissociation spectroscopy with free-electron lasers: on the road from small molecules to biomolecules. *Chem. A Eur. J.* 24, 3374–3390. doi: 10.1002/chem.201705692

Jones, D. R., Perttunen, C. D., and Stuckman, B. E. (1993). Lipschitzian optimization without the Lipschitz constant. *J. Optimization Theory Appl.* 79, 157–181. doi: 10.1007/BF00941892

Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallograph. Section A* 32, 922–923. doi: 10.1107/S0567739476001873

Kennedy, J., and Eberhart, R. (1995). Particle swarm optimization. 4, 1942–1948.

Kim, Y., Choi, S., and Kim, W. Y. (2014). Efficient basin-hopping sampling of reaction intermediates through molecular fragmentation and graph theory. *J. Chem. Theory Comput.* 10, 2419–2426. doi: 10.1021/ct500136x

Kohonen, T. (1990). The self-organizing map. *Proc. IEEE* 78, 1464–1480. doi: 10.1109/5.58325

Kong, X., Tsai, I.-A., Sabu, S., Han, C.-C., Lee, Y. T., Chang, H.-C., et al. (2006). Progressive stabilization of zwitterionic structures in [H(Ser)2–8]+ studied by infrared photodissociation spectroscopy. *Angewandte Chemie Int. Ed.* 45, 4130–4134. doi: 10.1002/anie.200600597

Krylov, E. V., Coy, S. L., and Nazarov, E. G. (2009). Temperature effects in differential mobility spectrometry. *Int. J. Mass Spectrometry* 279, 119–125. doi: 10.1016/j.ijms.2008.10.025

Krylov, E. V., and Nazarov, E. G. (2009). Electric field dependence of the ion mobility. *Int. J. Mass Spectrometry* 285, 149–156. doi: 10.1016/j.ijms.2009.05.009

Krylov, E. V., Nazarov, E. G., and Miller, R. A. (2007). Differential mobility spectrometer: model of operation. *Int. J. Mass Spectrom.* 266, 76–85. doi: 10.1016/j.ijms.2007.07.003

Kumar, A., and Zhang, K. Y. J. (2018). Advances in the development of shape similarity methods and their application in drug discovery. *Front. Chem.* 6, 1–21. doi: 10.3389/fchem.2018.00315

Lanucara, F., Holman, S. W., Gray, C. J., and Eyers, C. E. (2014). The power of ion mobility-mass spectrometry for structural characterization and the study of conformational dynamics. *Nat. Chem.* 6:281. doi: 10.1038/nchem.1889

Lapthorn, C., Pullen, F., and Chowdhry, B. Z. (2013). Ion mobility spectrometry-mass spectrometry (IMS-MS) of small molecules: separating and assigning structures to ions. *Mass Spectrometry Rev.* 32, 43–71. doi: 10.1002/mas.21349

Leary, R. H. (2000). Global optimization on funneling landscapes. *J. Global Optimiz.* 18, 367–383. doi: 10.1023/A:1026500301312

Lecours, M. J., Chow, W. C. T., and Hopkins, W. S. (2014). Density functional theory study of RhnS0,+/- and Rh-n+1(0,+/-) (n=1-9). *J. Phys. Chem. A* 118, 4278–4287. doi: 10.1021/jp412457m

Lemaire, J., Boissel, P., Heninger, M., Mauclaire, G., Bellec, G., Mestdagh, H., et al. (2002). Gas phase infrared spectroscopy of selectively prepared ions. *Phys. Rev. Lett.* 89:273002. doi: 10.1103/PhysRevLett.89.273002

Locatelli, M. (2005). On the multilevel structure of global optimization problems. *Comput. Optimization Appl.* 30, 5–22. doi: 10.1007/s10589-005-4561-y

Locatelli, M., and Schoen, F. (2013). Global optimization: theory, algorithms, and applications. *Soc. Industr. Appl. Mathematics*. 30, 5–22. doi: 10.1137/1.9781611972672

Ma, L., Ren, J., Feng, R., Zhang, K., and Kong, X. (2018). Structural characterizations of protonated homodimers of amino acids: revealed by infrared multiple photon dissociation (IRMPD) spectroscopy and theoretical calculations. *Chin. Chem. Lett.* 29, 1333–1339. doi: 10.1016/j.cclet.2018.02.008

Macaleese, L., and Maître, P. (2007). Infrared spectroscopy of organometallic ions in the gas phase: from model to real world complexes. *Mass Spectrometry Rev.* 26, 583–605. doi: 10.1002/mas.20138

Maeda, S., Taketsugu, T., and Morokuma, K. (2014). Exploring transition state structures for intramolecular pathways by the artificial force induced reaction method. *J. Computational Chem.* 35, 166–173. doi: 10.1002/jcc.23481

Martinez, T., and Schulten, K. (1991). *Artificial Neural Networks, Vol. 1*, eds T. Kohonen, K. Kakisara, O. Simula, J. Kangas (Amsterdam: Elsevier B.V.), 397–402.

Mason, E. A., and Mcdaniel, E. W. (1988). *Transport Properties of Ions in Gases.* New York, NY: John Wiley and Sons.

Michener, C. D., and Sokal, R. R. (1957). A quantitative approach to a problem in classification. *Evolution* 11, 130–162. doi: 10.1111/j.1558-5646.1957.tb02884.x

Mino, W. K., Gulyuz, K., Wang, D., Stedwell, C. N., and Polfer, N. C. (2011). Gas-phase structure and dissociation chemistry of protonated tryptophan elucidated by infrared multiple-photon dissociation spectroscopy. *J. Phys. Chem. Lett.* 2, 299–304. doi: 10.1021/jz1017174

Montavon, G., Hansen, K., Fazli, S., Rupp, M., Biegler, F., Ziehe, A., et al. (2012). "Learning invariant representations of molecules for atomization energy prediction," in Advances in Neural Information Processing Systems 25, eds F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Curran Associates, Inc.), 440–448.

Nanita, S. C., and Cooks, R. G. (2006). Serine octamers: Cluster formation, reactions, and implications for biomolecule homochirality. *Angew. Chem. Int. Ed.* 45, 554–569. doi: 10.1002/anie.200501328

Noguera, M., RodríGuez-Santiago, L., Sodupe, M., and Bertran, J. (2001). Protonation of glycine, serine and cysteine. Conformations, proton affinities and intrinsic basicities. *J. Mol. Struc. Theochem*. 537, 307–318. doi: 10.1016/S0166-1280(00)00686-2

Oh, H. B., Lin, C., Hwang, H. Y., Zhai, H., Breuker, K., Zabrouskov, V., et al. (2005). Infrared photodissociation spectroscopy of electrosprayed ions in a Fourier transform mass spectrometer. *J. Am. Chem. Soc.* 127, 4076–4083. doi: 10.1021/ja040136n

Oh, K. J., and Zeng, X. C. (1999). Formation free energy of clusters in vapor-liquid nucleation: a Monte Carlo simulation study. *J. Chem. Phys.* 110, 4471–4476. doi: 10.1063/1.478331

Olson, B., Hashmi, I., Molloy, K., and Shehu, A. (2012). Basin hopping as a general and versatile optimization framework for the characterization of biological macromolecules. *Adv. Artificial Intellig.* 2012:19. doi: 10.1155/2012/674832

Oomens, J., Steill, J. D., and Redlich, B. (2009). Gas-phase IR spectroscopy of deprotonated amino acids. *J. Am. Chem. Soc.* 131, 4310–4319. doi: 10.1021/ja807615v

Paglia, G., and Astarita, G. (2017). Metabolomics and lipidomics using traveling-wave ion mobility mass spectrometry. *Nat. Protoc.* 12:797. doi: 10.1038/nprot.2017.013

Parneix, P., Basire, M., and Calvo, F. (2013). Accurate modeling of infrared multiple photon dissociation spectra: the dynamical role of anharmonicities. *J. Phys. Chem. A* 117, 3954–3959. doi: 10.1021/jp402459f

Peng, C., Ayala, P. Y., Schlegel, H. B., and Frisch, M. J. (1996). Using redundant internal coordinates to optimize equilibrium geometries and transition states. *J. Comput. Chem.* 17, 49–56. doi: 10.1002/(SICI)1096-987X(19960115)17:1<49::AID-JCC5>3.0.CO;2-0

Peng, C., and Bernhard Schlegel, H. (1993). Combining synchronous transit and quasi-newton methods to find transition states. *Israel J. Chem.* 33, 449–454. doi: 10.1002/ijch.199300051

Piela, L., Olszewski, K. A., and Pillardy, J. (1994). On the stability of conformers. *Theochem. J. Mol. Struc.* 114, 229–239. doi: 10.1016/0166-1280(94)80105-3

Polfer, N. C. (2011). Infrared multiple photon dissociation spectroscopy of trapped ions. *Chem. Soc. Rev.* 40, 2211–2221. doi: 10.1039/c0cs00171f

Rahman, S. A., Bashton, M., Holliday, G. L., Schrader, R., and Thornton, J. M. (2009). Small Molecule Subgraph Detector (SMSD) toolkit. *J. Cheminform.* 1, 1–13. doi: 10.1186/1758-2946-1-12

Robinson, E. W., Shvartsburg, A. A., Tang, K., and Smith, R. D. (2008). Control of ion distortion in field asymmetric waveform ion mobility spectrometry via variation of dispersion field and gas temperature. *Analyt. Chem.* 80, 7508–7515. doi: 10.1021/ac800655d

Röder, K., and Wales, D. J. (2018). Mutational basin-hopping: combined structure and sequence optimization for biomolecules. *J. Phys. Chem. Lett.* 9, 6169–6173. doi: 10.1021/acs.jpclett.8b02839

Scheraga, H. A. (1992). Some approaches to the multiple-minima problem in the calculation of polypeptide and protein structures. *Int. J. Quantum Chem.* 42, 1529–1536. doi: 10.1002/qua.560420526

Schmidt, J., Meyer, M. M., Spector, I., and Kass, S. R. (2011). Infrared multiphoton dissociation spectroscopy study of protonated p-aminobenzoic acid: does electrospray ionization afford the amino- or carboxy-protonated ion? *J. Phys. Chem. A* 115, 7625–7632. doi: 10.1021/jp203829z

Schofield, D. P., Kjaergaard, H. G., Matthews, J., and Sinha, A. (2005). The OH-stretching and OOH-bending overtone spectrum of HOONO. *J. Chem. Phys.* 123:134318. doi: 10.1063/1.2047574

Scutelnic, V., Perez, M. A. S., Marianski, M., Warnke, S., Gregor, A., Seo, J., et al. (2018). The structure of the protonated serine octamer. *J. Am. Chem. Soc.* 140, 7554–7560. doi: 10.1021/jacs.8b02118

Seo, J., Hoffmann, W., Malerz, S., Warnke, S., Bowers, M. T., Pagel, K., et al. (2018). Side-chain effects on the structures of protonated amino acid dimers: a gas-phase infrared spectroscopy study. *Int. J. Mass Spectrometry* 429, 115–120. doi: 10.1016/j.ijms.2017.06.011

Seo, J., Warnke, S., Pagel, K., and Bowers, M. T. A. (2017). Infrared spectrum and structure of the homochiral serine octamer-dichloride complex. *Nat. Chem.* 9, 1263–1268. doi: 10.1038/nchem.2821

Shi, L. T., Wang, Z. Q., Hu, C. E., Cheng, Y., Zhu, J., and Ji, G. F. (2019). Possible lower energy isomer of carbon clusters C n (n = 11, 12) via particle swarm optimization algorithm: Ab initio investigation. *Chem. Phys. Lett.* 721, 74–85. doi: 10.1016/j.cplett.2019.02.028

Siems, W. F., Viehland, L. A., and Hill, H. H. (2016). Correcting the fundamental ion mobility equation for field effects. *Analyst* 141, 6396–6407. doi: 10.1039/C6AN01353H

Sokal, R. R., and Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.* 38, 1409–1438.

Stedwell, C. N., Galindo, J. F., Gulyuz, K., Roitberg, A. E., and Polfer, N. C. (2013). Crown complexation of protonated amino acids: Influence on IRMPD spectra. *J. Phys. Chem. A* 117, 1181–1188. doi: 10.1021/jp305263b

Steill, J. D., Szczepanski, J., Oomens, J., Eyler, J. R., and Brajter-Toth, A. (2011). Structural characterization by infrared multiple photon dissociation spectroscopy of protonated gas-phase ions obtained by electrospray ionization of cysteine and dopamine. *Anal. Bioanal. Chem.* 399, 2463–2473. doi: 10.1007/s00216-010-4582-y

Storn, R., and Price, K. (1997). Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *J. Global Optimiz.* 11, 341–359. doi: 10.1023/A:1008202821328

Sunahori, F. X., Yang, G., Kitova, E. N., Klassen, J. S., and Xu, Y. (2013). Chirality recognition of the protonated serine dimer and octamer by infrared multiphoton dissociation spectroscopy. *Phys. Chem. Chem. Phys.* 15, 1873–1886. doi: 10.1039/C2CP43296J

Tian, Z., and Kass, S. R. (2009). Gas-Phase versus Liquid-Phase Structures by Electrospray Ionization Mass Spectrometry. *Angew. Chem. Int. Ed.* 48, 1321–1323. doi: 10.1002/anie.200805392

Tian, Z. X., and Kass, S. R. (2008). Does Electrospray ionization produce gas-phase or liquid-phase structures? *J. Am. Chem. Soc.* 130:10842. doi: 10.1021/ja802088u

Vehkamäki, H. (2006). *Classical Nucleation Theory in Multicomponent Systems.* Berlin; Heidelberg; New York, NY: Springer-Verlag.

Verkouteren, J. R., and Staymates, J. L. (2011). Reliability of ion mobility spectrometry for qualitative analysis of complex, multicomponent illicit drug samples. *Forensic Sci. Int.* 206, 190–196. doi: 10.1016/j.forsciint.2010.08.005

Viehland, L. A., and Mason, E. A. (1995). Transport properties of gaseous-ions over a wide energy-range.4 *Atomic Data Nuclear Data Tables* 60, 37–95. doi: 10.1006/adnd.1995.1004

Wales, D. J., and Doye, J. P. K. (1997). Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *J. Phys. Chem. A* 101, 5111–5116. doi: 10.1021/jp970984n

Wales, D. J., Miller, M. A., and Walsh, T. R. (1998). Archetypal energy landscapes. *Nature* 394, 758–760. doi: 10.1038/29487

Wales, D. J., and Scheraga, H. A. (1999). Review: chemistry - global optimization of clusters, crystals, and biomolecules. *Science* 285, 1368–1372. doi: 10.1126/science.285.5432.1368

Wang, J., Wang, W., Kollman, P. A., and Case, D. A. (2006). Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* 25, 247–260. doi: 10.1016/j.jmgm.2005.12.005

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* 58, 236–244. doi: 10.1080/01621459.1963.10500845

Weininger, D. (1988). SMILES, a chemical language and information system: 1: introduction to methodology and encoding rules. *J. Chem. Inform. Comp. Sci.* 28, 31–36. doi: 10.1021/ci00057a005

Wickelmaier, F. (2003). *An Introduction to MDS*. Aalborg: Aalborg Universitetsforlag.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 14 August 2019
doi: 10.3389/fchem.2019.00576

Specific Reaction Parameter Multigrid POTFIT (SRP-MGPF): Automatic Generation of Sum-of-Products Form Potential Energy Surfaces for Quantum Dynamical Calculations

Ramón L. Panadés-Barrueta[1], *Emilio Martínez-Núñez*[2] *and Daniel Peláez*[1]*

[1] *Laboratoire de Physique des Lasers, Atomes et Molécules (PhLAM), Université de Lille, Villeneuve-d'Ascq, France,*
[2] *Departamento de Química Física, Facultade de Química, Universidade de Santiago de Compostela, Santiago de Compostela, Spain*

OPEN ACCESS

Edited by:
Nino Russo,
University of Calabria, Italy

Reviewed by:
Hans-Dieter Meyer,
Heidelberg University, Germany
Riccardo Conte,
University of Milan, Italy

***Correspondence:**
Daniel Peláez
daniel.pelaez-ruiz@univ-lille.fr

Specialty section:
This article was submitted to Theoretical and Computational Chemistry,
a section of the journal Frontiers in Chemistry

Received: *04 June 2019*
Accepted: *30 July 2019*
Published: *14 August 2019*

Citation:
Panadés-Barrueta RL, Martínez-Núñez E and Peláez D (2019) Specific Reaction Parameter Multigrid POTFIT (SRP-MGPF): Automatic Generation of Sum-of-Products Form Potential Energy Surfaces for Quantum Dynamical Calculations. Front. Chem. 7:576. doi: 10.3389/fchem.2019.00576

We present Specific Reaction Parameter Multigrid POTFIT (SRP-MGPF), an automated methodology for the generation of global potential energy surfaces (PES), molecular properties surfaces, e.g., dipole, polarizabilities, etc. using a single random geometry as input. The SRP-MGPF workflow integrates: (i) a fully automated procedure for the global topographical characterization of a (intermolecular) PES based on the Transition State Search Using Chemical Dynamical Simulations (TSSCDS) family of methods;i (ii) the global optimization of the parameters of a semiempirical Hamiltonian in order to reproduce a given level of electronic structure theory; and (iii) a tensor decomposition algorithm which turns the resulting SRP-PES into sum of products (Tucker) form with the Multigrid POTFIT algorithm. The latter is necessary for quantum dynamical studies within the Multiconfiguration Time-Dependent Hartree (MCTDH) quantum dynamics method. To demonstrate our approach, we have applied our methodology to the *cis-trans* isomerization reaction in HONO in full dimensionality (6D). The resulting SRP-PES has been validated through the computation of classical on-the-fly dynamical calculations as well as calculations of the lowest vibrational eigenstates of HONO as well as high-energy wavepacket propagations.

Keywords: PES, sums-of-products, tensor-decomposition, quantum dynamics, reparametrized semiempirical, TSSCDS, global optimization

1. INTRODUCTION

A detailed knowledge of the topography of a Potential Energy Surface (PES) is a highly desirable prerequisite for the simulation of any dynamical process. Topography on its own, however, does not fully determine the behavior of a system and dynamics calculations become mandatory (Tuckerman et al., 2002; Peláez et al., 2014). Furthermore, for an accurate theoretical description of molecular processes (spectroscopy, reactivity), one should, if possible, resort to nuclear quantum dynamics calculations (Gatti, 2014). In the specific case of vibrational problems, powerful methods based on the resolution of the time-independent Schrödinger equation exist such as vibrational self-consistent field/vibrational configuration interaction (VSCF/VCI) (Rauhut, 2007; Neff and Rauhut, 2009), vibrational second-order perturbation theory (VPT2) (Barone, 2005) and vibrational coupled-cluster theory (Christiansen, 2004). For an extensive and recent review of some of them, the reader is referred to a recent publication (Puzzarini et al., 2019). However, owing to our interest in describing chemical processes, we shall turn our attention toward methods able to describe wave packet propagations. In this context, within the last few years, we have experienced a boost in dynamical methodologies capable of describing the dynamics of molecular systems up to medium-large size, ranging from semiclassical (Levine et al., 2008; Shalashilin, 2010) to fully quantal (Gatti, 2014). With respect to the latter, by far, the most popular approaches nowadays are those based on, or related to, the grid-based Multiconfiguration Time-Dependent Hartree (MCTDH) algorithm (Beck et al., 2000). In MCTDH, a molecular wavefunction (WF) is expanded in a basis of time-dependent nuclear orbitals. Taken MCTDH as reference, two powerful multiconfigurational methods exist. On the one hand, the partially grid-based G-MCTDH method in which some of the time-dependent basis functions are substituted by (typically frozen) Gaussians functions (G) (Burghardt et al., 2008), and the Variational Multiconfigurational Gaussian (vMCG) method (Richings et al., 2015) (and its direct-dynamics (DD) extension) which are grid-free and use Gaussian functions only. For the sake of completeness, one should mention the recent and promising direct-dynamics approach of MCTDH by Richings and Habershon (2018).

It should be evident that the quality of the results of any dynamical calculation is limited by the accuracy and efficiency of the underlying electronic structure method used to represent the PES, either globally (as in grid-based methods) or locally (on-the-fly approaches). When expressed *globally* on a grid, formally as a multidimensional tensor, the limitation lies on the number of dynamical degrees of freedom and the possibility of fitting the PES to an appropriate functional form. In the case of on-the-fly methods, on the other hand, the number of degrees of freedom (DOF) it is not the main limiting factor but the *number of electrons*, in other words, the level of theory and its performance in the form of electronic structure software *calls* (energies, gradients, Hessians) at each time-step. This fact constrains on-the-fly approaches to modest levels of theory.

Obtaining a fit for a high-dimensional PES is a complex and tedious task. Whatever the approach, any fitting procedure requires a more or less large set of reference values (molecular energies and/or gradients and, possibly, properties such as dipoles) which will constitute the data to which an algorithm will try to fit a given function. *Ad hoc* analytical functions are usually added to the resulting fit in order to ensure a correct physical behavior, for instance in the asymptotic regions, or to guarantee a correct periodicity of the potential as in the case of rotors. Focusing on the fitting methods typically used in combination with nuclear quantum dynamical approaches, many techniques have been proposed. To name but a few, popular methods include the permutationally invariant polynomials (Braams and Bowman, 2009), the interpolating moving least-squares (Dawes et al., 2007), the triatomics-in-molecules approximation (Sanz-Sanz et al., 2013), Shephard interpolation schemes (Frankcombe and Collins, 2011). Moreover, for more than a decade now, Neural Network (NN) approaches have (re)gained preeminence being triggered by the pioneering work of Manzhos and Carrington (2006) and, very recently, their application to MCTDH by Pradhan and Brown (2017). In this line, Jiang and Guo have gone a step further and have developed a NN approach with implicit nuclear permutational symmetry (Jiang and Guo, 2014). For the sake of completeness, one should mention the works of Rauhut (2004) and Sparta et al. (2010) in which PESs for vibrational calculations are generated in an automated and adaptive fashion. Powerful and accurate as these methods are, a high degree of expertise is still required to master and to apply these techniques, particularly for medium-large systems (≥6D), thus preventing them from a wider-spread use. Furthermore, in studies where external fields (e.g., a laser) are needed, surfaces of molecular properties are also required and, as a consequence, extra fits are necessary.

In this work, we present Specific Reaction Parameter Multigrid POTFIT (SRP-MGPF), a method which provides a well-balanced solution to the aforementioned issues. SRP-MGPF is able to generate a chemically-accurate PES as well as the same-level-of-theory molecular properties surfaces, starting from a single input geometry and requiring minimal intervention of the user. In this sense, we can safely affirm that the procedure is *quasi* black-box in nature. SRP-MGPF relies on three main steps: (i) generation of a set of reference geometries (energies and properties); (ii) reparametrization of a semiempirical Hamiltonian (Specific Reaction Parameter Hamiltonian, SRP) based on the previous information; and (iii) tensor-decomposing the SRP with MGPF. We shall focus on the *standard* MCTDH method for which a global PES needs to be fitted into some kind of functional form and, typically, refitted to a grid. Furthermore, our results can also be directly applied to any on-the-fly methodology. It should be highlighted at this point that reparametrized semiempirical Hamiltonians have been typically used in direct dynamics studies as well as in kinetic studies (Rossi and Truhlar, 1995; Troya, 2005; Rodríguez-Fernández et al., 2017). Moreover, semiempirical Hamiltonians have been successfully used in describing dynamics on electronically excited states (Toniolo et al., 2003; Silva-Junior and Thiel, 2010). It should be stressed that SRP methods qualify as quantum chemical ones. As such an SRP does not include, necessarily, any fitting functions. Hence, the SRP parameters obtained through our fitting process will define a level of electronic structure close to a high-level reference one.

In our approach, as generator set for the reference fitting points, we employ the so-called Reaction Network (RXN) (Martínez-Núñez, 2015b), i.e., the *complete* set of stationary points (minima, transition states,…) of a PES. The RXN captures the main topographical (even topological) features of the target PES and thus constitutes a sensible choice for the reference set. Characterization of the topography of a PES is, however, not an evident task. To this end, we make use of the recently developed Transition State Search Using Chemical Dynamics Simulations

(TSSCDS) (Martínez-Núñez, 2015a,b) method which relies on the efficient sampling of configuration space combined with a graph-theory based identification of transition state (TS) structures, which are finally optimized and the corresponding Minimum Energy Paths obtained with standard methods. The TSSCDS approach has been recently extended to specifically study van der Waals complexes (vdW) or, more generally, non-covalently bound systems (vdW-TSSCDS) (Kopec et al., 2019).

A set of *optimal* semiempirical Hamiltonian parameters is then obtained by global minimization of the Root-Mean Square Error (RMSE) between a set of reference *ab initio* energies, for instance, (on the RXN-derived geometries) and the corresponding SRP ones. The SRP approach to PESs presents interesting features that make it very appealing when compared to formally higher-level methods (Density Functional Theory, DFT, or *ab initio*). First, SRPs are fast-computing parametrized electronic structure methods, some of the integrals are neglected while the remaining are parametrized to reproduce high-level results. As such, they typically exhibit a *correct* physical behavior. Second, in contrast to other *fitting* procedures (for instance based on any kind of polynomial expansions or neural networks), SRPs exhibit a correct behavior outside the fitting boundaries, *if* the SRP parameters remain somewhat physical (*small* variation with respect to their reference values). Third, by varying the SRP parameters we can simultaneously fit both energies and the molecular properties accessible to the semiempirical software. It should be highlighted that in the usual approach energies and properties (e.g., dipole) are computed at a set of reference geometries and then need to be independently fitted to either potential energy surfaces or property surfaces (x-dipole, y-dipole, etc.). In contrast, in our method a single optimization process suffices to yield a simultaneous fit of all properties simultaneously, provided that information on the desired properties is included in the reference data. Last, but not least, the number of parameters is *independent* of the number of atoms. They only depend on the number of different atoms (and possibly on their chemical function) and, as such, it is in a sense not affected by the curse of dimensionality. In our specific approach, we have used as base model chemistry the Parametric Method 7 (PM7) method as implemented in the OpenMOPAC software package (Stewart, 2016). This choice is justified by the quality of the obtained results as well as its efficiency in terms of computational time (PM7 is orders of magnitude faster than *ab initio* methods) (Stewart, 2013).

The final step, specific for grid-based methods, is the tensor-decomposition of the SRP-PES into an appropriate form. To this end, we utilize the Multigrid POTFIT (MGPF) algorithm (Peláez and Meyer, 2013), succinctly described in section 2.3. MGPF has been successfully applied to the computation of vibrational eigenstates (Peláez et al., 2014), infrared (IR) spectra (Peláez and Meyer, 2017), and electron dynamics including continuum (Haller et al., 2019) With SRP-MGPF, owing to the extreme efficiency of the semiempirical calculations, we can directly generate the SRP-PES on a grid.

This manuscript is structured as follows. In section 2 we provide a succinct introduction to the methods employed in our workflow. In section 3, which presents the application of our novel methodology to the HONO molecule in full-dimensionality, we carefully discuss all specific aspects related to the actual calculations. Section 4 concludes the paper and gives some hints on future developments and possible applications of the method.

2. THEORY AND COMPUTATIONAL DETAILS

Our automated methodology for computing a global PES consists of three steps: (1) automatic and global determination of stationary points (minima and transition states), as well as the corresponding Intrinsic Reaction Coordinate paths (IRCs), the so-called Reaction Network (RXN); (2) reparametrization of a semiempirical Hamiltonian (SRP) to reproduce a desired level of electronic structure theory (e.g., *ab initio*) using the RXN and neighboring points; and (3) tensor-decomposition of the SRP Hamiltonian with the MGPF algorithm. It should be noted that after stage (2), we already have a global PES which can be used in conjunction with any type of *on-the-fly* dynamics scheme. We shall describe in the following each of the above mentioned stages. First of all, we shall discuss our specific procedure for the reparametrization of semiempirical Hamiltonians. Then, we shall present our way of generating a set of reference points based on the RXN obtained using the (vdW-)TSSCDS method (Martínez-Núñez, 2015a,b). Subsequently, we shall discuss how we integrate this information in combination with the NLOpt (Johnson, 2011) library and the openMOPAC software (Stewart, 2016) to produce an *optimal* set of SRP parameters. The resulting SRP-PES is then *interfaced* with MCTDH through the Multigrid POTFIT program (Peláez and Meyer, 2013) thus generating a SRP-MGPF PES on the grid and in sums-of-product (SOP) form.

Finally, it should be highlighted that, for the graphical representations, we have made extensive use of the SciPy scientific tools by Jones et al. (2001).

2.1. Global Optimization of Semi-empirical Hamiltonians Parameters

Semiempirical potentials can be seen as parametrized Hartree-Fock methods in which some of the electronic integrals are either neglected or replaced by parameters obtained as fitting constants using large sets of reference data (high-level *ab initio* calculations and/or experimental data) (Stewart, 2013; Thiel, 2014). In this sense, semiempirical methods lie somewhere between force fields and *ab initio* methods (Stewart, 2013). Owing to the lower amount of integral calculations, semiempirical methods are orders of magnitude faster than *ab initio* methods and, hence, they are routinely used in the study of large systems (Christensen et al., 2016). In addition to this, with a suitable configuration interaction formalism, semiempirical methods can also be used for the study of excited states (Toniolo et al., 2003; Silva-Junior and Thiel, 2010). A milestone in the usage of semiempiricals was achieved by Rossi and Truhlar (1995) who introduced the idea of reparametrizing a semiempirical Hamiltonian in order to reproduce a given high-level *ab initio* level of theory for a *specific* chemical reaction (or family thereof), hence the name of

Specific Reaction Parameter (SRP) Hamiltonians. Since then, this technique has been successfully applied to the study of chemical reactions of large-dimensional systems using classical dynamics (Layfield et al., 2008) as well as to kinetic studies (Rodríguez-Fernández et al., 2017). In the present work, we go a step further and will use the SRP approach for the generation of a PES suitable for quantum dynamical studies. To this end, we used the publicly available non-linear global optimization library NLOpt (Johnson, 2011) to reparametrize the PM7 semiempirical model (Stewart, 2013) as implemented in openMOPAC (Stewart, 2016). The choice of PM7 responds not only to its proven accuracy but also to the fact that it includes diatomic parameters in addition to the standard atomic ones, thus providing extra flexibility to the optimization process (Stewart, 2013). Hereafter, we shall refer to the set of SRP parameters as $\{\zeta_i\}_{i=1}^{D}$, being D the total number of parameters. It is important to notice that the latter depends on the number of *atom types* and *not* on the dimensionality of the system. It should be stressed that we are dealing with a *fitting function* which has an implicit physical character (HF-like) and, as such, it is expected to yield a *global* qualitatively-correct behavior and to require less fitting points than other traditional fitting approaches.

The problem that concerns us is thus the global optimization of a deterministic non-linear objective function $\chi(\boldsymbol{\zeta})$: $\mathbb{R}^D \rightarrow \mathbb{R}$, Equation (1), with a bounded parameter space ($\zeta_i \in [\zeta_i^{min}, \zeta_i^{max}], i = 1, \ldots, D$). In our specific case, we do not make use of the derivatives of this target function since: (i) the analytical expressions are unavailable; (ii) their numerical determination would be expensive and, more importantly, complicated due to the highly-corrugated character of the RMSE landscape (see **Figure 1**). We shall consider then a *derivative-free optimization* algorithm (Rios and Sahinidis, 2013). As general expression of the objective function (χ) we have considered a *rms-like* function (see Equation 1) composed by two terms: (i) a first one accounting for the error in the energies and (ii) a the second one corresponding to the error in the harmonic frequencies of the stationary points of the PES, with respect to our reference calculations. We have observed that the inclusion of the latter helps to preserve the correct topography of the PES, for instance the first order saddle point character of transition states.

$$\chi_0(\boldsymbol{\zeta}) = \sqrt{\sum_{i=1}^{n} \frac{\omega_E(E_i^{ab}) \cdot [E_i^{ab} - E_i^{srp}(\boldsymbol{\zeta})]^2}{n} + \sum_{j=1}^{m} \frac{\omega_F(\Delta F_j) \cdot [F_j^{ab} - F_j^{srp}(\boldsymbol{\zeta})]^2}{m}} \tag{1}$$

where $\boldsymbol{\zeta}$ is a vector containing the semiempirical parameters and n, m represent the number of (relative) energy data points ($E^{ab/srp}$) and harmonic frequencies ($F^{ab/srp}$), respectively, the labels referring to *ab initio* (*ab*) and semiempirical (*srp*) data. The weighting functions $\omega_E(E_i^{ab})$ and $\omega_F(\Delta F_j)$ (with $\Delta F_j = F_j^{ab} - F_j^{srp}$) have been defined as exponential step functions:

$$f(x) = \begin{cases} 1 & x \leq \alpha \\ e^{\beta(x-\alpha)} & x > \alpha \end{cases} \tag{2}$$

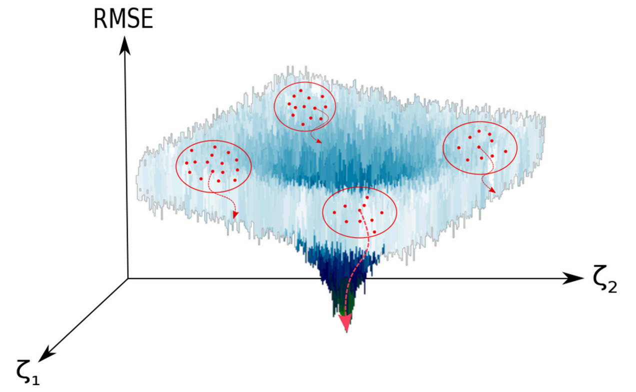


FIGURE 1 | Graphical representation of the optimization process of the set of SRP parameters ($\{\boldsymbol{\zeta}\}$). The vertical axis displays the RMSE between our reference data and our target function (see Equation 3), which in the figure depends just on two parameters (ζ_1, ζ_2). Non-overlapping clusters (red dots enclosed in a red circle) of walkers (red dots) are generated. In each cluster, the *optimal* solution is locally minimized (red dotted curved arrows) and compared to the rest of solutions. For a large enough number of clusters, convergence to the global minimum is guaranteed. In this representation, we have used a modified Ackley function (Ackley, 1987).

where α, β are parameters adjusted *a priori* and x corresponds to the selected argument ($E_i^{ab}, \Delta F_j$). However, in practice, we have obtained satisfactory results with a much simpler expression:

$$\chi_1(\boldsymbol{\zeta}) = \sqrt{\sum_{i=1}^{n+m} \frac{\omega_G(G_i^{ab}) \cdot [G_i^{ab} - G_i^{srp}(\boldsymbol{\zeta})]^2}{n+m}} \tag{3}$$

where $G_i = E_i || F_i$ are the components of a vector constructed by concatenating the vectors of energies and harmonic frequencies, respectively. As strategy, we have performed a global optimization step followed by local optimizations in order to refine the results. For the former, we used the Multi-Level Single-Linkage (MLSL) algorithm (Kan and Timmer, 1987) and for the latter we used the Bound Optimization BY Quadratic Approximation (BOBYQA) (Powell, 2009).

2.2. Automated Generation of the Set of Reference Points

In the following, we shall describe our automated methodology for the generation of a set of fitting points for the reparametrization of a semiempirical Hamiltonian. In brief, we propose the use of the *whole* set of stationary points of a given PES, the so-called RXN (Martínez-Núñez, 2015a,b; Kopec et al., 2019), as initial set from which neighboring geometries spanning the region of configuration space of interest will be generated. The main advantage of our method is that starting from a *single* initial input geometry, a *global* Potential Energy Surface is generated.

We propose as first step the determination of the ensemble of stationary points (RXN) on a given PES which will be used as seed for the subsequent generation of the remaining

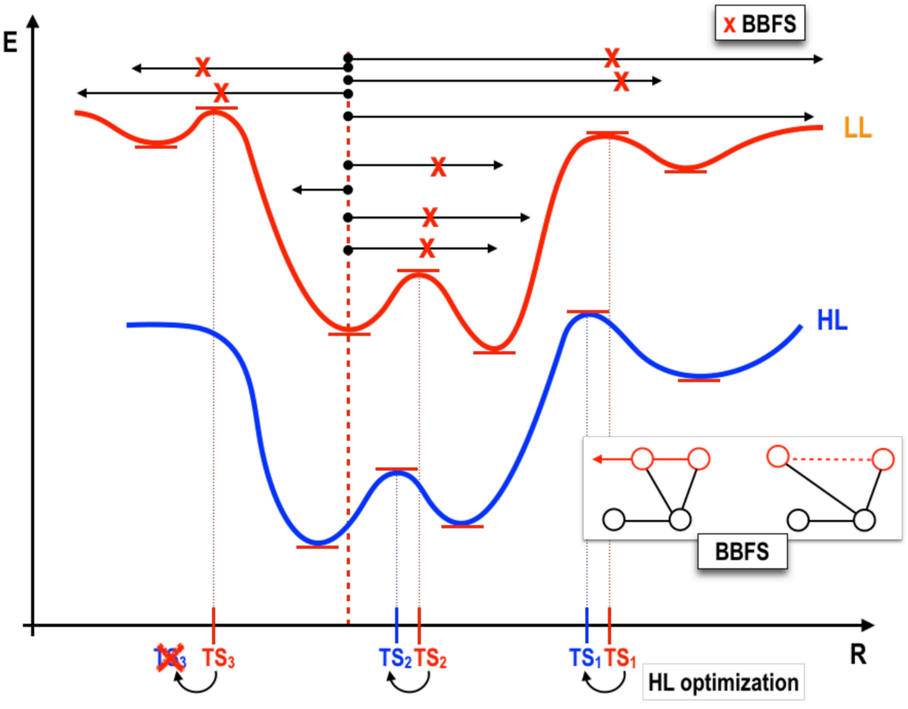


FIGURE 2 | One-dimensional representation of the TSSCDS procedure. A low level (LL) PES (upper energy curve, in red) is sampled starting from a given minimum (geometry indicated by a red dotted line). Classical random trajectories (black arrows) in combination with a graph theory based method (Bond Breaking/Formation Search, BBFS Martínez-Núñez, 2015a,b) lead to the determination of TS candidate structures (marked as x in red bold font), compatible with the total energy of the trajectories, from which LL optimizations are started. Subsequent optimization at the desired high-level (HL) are performed using the LL TS as guess structures.

fitting points. Indeed, the stationary points correspond to the molecular configurations which carry the most relevant topographical information of a given PES and, as such, make ideal candidates for fitting purposes. Finding stationary points, however, is a very tedious task which heavily relies on large amounts of chemical intuition. Fortunately, a family of methods for the automated determination of the RXN has been recently proposed, the so-called Transition State Search Using Chemical Dynamics (TSSCDS) (Martínez-Núñez, 2015a,b) as well as its generalization, vdW-TSSCDS (Kopec et al., 2019). The former is optimal for the study of unimolecular processes whereas the latter has been specifically designed to study non-covalently bound systems. The workflow in both cases is analogous (see **Figure 2**) and the difference lies in the way transition states (TSs) are characterized. Starting from an initial random geometry (or small set thereof), a large number of high-energy classical trajectories is run using a low-level (LL) of electronic structure theory (semiempirical in our case, other methods are also possible) to compute the forces. The geometries along these trajectories are analyzed by a graph-theory based algorithm (Bond Breaking/Formation Search, BBFS Martínez-Núñez, 2015a,b; Kopec et al., 2019) which detects conformations in which bonds are broken and/or formed. It should be highlighted that this step is precisely what determines the difference between TSSCDS and vdW-TSSCDS. In the former, a square connectivity matrix based on covalent distances is defined, whereas in the latter this matrix takes block-diagonal form and includes both covalent and non-covalent (van der Waals) distances, thus allowing for the determination of non-covalent saddle points. The so-determined structures, candidates to TSs, are optimized at the LL and subsequently reoptimized at an appropriate higher level of theory, say, *ab initio* or DFT. Obviously, this process can be continued by further refinements. From this set of final high-level TSs, IRC calculations connecting minima are performed. And, as a result of this, the so-called Reaction Network (RXN) is obtained, that is, all stationary structures together with their connectivities compatible with a given total energy (that of the initial classical trajectories). For further details on the method, the interested reader is referred to the original publications (Martínez-Núñez, 2015a,b; Kopec et al., 2019). As indicated, the RXN will serve us as initial set from which the full set of fitting points will be generated. The total number of stationary points (N_{RXN}) is:

$$N_{RXN} = n_{min} + n_{TS} + n_{asymp} + \dots, \tag{4}$$

where n_X, (with X=min, TS, asymp,...) is the number of minima, transition states (TS), asymptotic products, respectively. This initial set will be extended by systematically adding a set of *neighboring* geometries. This can be achieved in different ways. In our case, we have chosen to distort each of the N_{RXN} points following an n-body type of scheme inspired by a previous work (Pradhan and Brown, 2017). The novelty of our procedure lies in the fact that we observe convergence in the RMSE at each order of the expansion. As it will be clear later, this convergence provides us with an efficient error control and allows to determine a minimal number of fitting points necessary to achieve a given RMSE. The total number of fitting points (N_{ref}) can be calculated as:

$$N_{ref} = N_{RXN} \cdot [\sum_{i\in 1D}^{f} N_i^{(1D)} + \sum_{i\in 2D}^{f} N_i^{(2D)} + \ldots] + rnd(fD) \\ + \sum_{i}^{n_{TS}} N_i^{IRC} + \sum_{i}^{n_{asymp}} N^{(asymp)} + \ldots \quad (5)$$

where f is the number of degrees of freedom of the molecular system, N is the number of generated reference geometries of a given type, for instance, N^{nD} are grid points from a n-dimensional (D) grid and N_i^{IRC} are the IRC points stemming from TS_i, $rnd(fD)$ are random geometries in the full-D configuration space, n is the number of stationary points of a kind. Considering, for instance, a normal mode or internal coordinate local representation, 1D would refer to displacements along each mode/coordinate (leaving the remaining coordinates fixed at their equilibrium values) and nD refers to grids of points generated through simultaneous displacement along n modes/coordinates, leaving the remaining fixed as before.

Our goal is now to determine the minimum number of fitting points leading to the smallest possible RMSE (defined as the difference between reference PES and SRP-PES), or, in other words, the *optimal* set of SRP parameters ($\{\zeta_{opt}\}$). It should be emphasized that we are dealing with moderate-size configuration spaces, in our specific case HONO (6D), the parameter space is 34-dimensional. Hence, in order to systematically search for the global minimum in SRP parameter space ($\{\zeta\}$), we increase the number of reference points in a controlled way according to the following prescription. Starting with the PM7 parameters ($\{\zeta_{PM7}\}$) as initial guess, the RMSE($\boldsymbol{\zeta}$) landscape is explored in a first stage using a small number of ab initio reference data and a big number of iterations (typically of the order of 10^5) of the non-linear optimization algorithm (MLSL in our case). This allows to locate the most-likely candidate parameter set to global minimum. The latter is used as a guess in subsequent local optimization stages (BOBYQA). At each of these, extended sets of points are generated in the form of *nD* distortions. At each level (1D, 2D, etc.) and for each set, we carry out local optimizations, compare the resulting RMSEs and take as *optimal* the number of points (set) that leads to a satisfactory value of RMSE, in the form of convergence, thus guaranteeing the condition of minimum number of points.

2.3. Generation of the SRP-MGPF Potential Energy Surface

As any other grid-based method, MCTDH quantum dynamics relies on a discretization of the configuration space known as *primitive* grid (Kosloff, 1988). In an f-dimensional molecular system (typically f = 3N-6, with N being the number of atoms), a set of $i_\kappa = 1, \ldots, N_\kappa$ grids points is defined for the κ-th DOF with $\kappa = 1, \ldots, f$. In other words, a given grid point $I \equiv (i_1, \ldots, i_f)$ has an associated molecular configuration ($Q \equiv (q_i, \ldots, q_f)$). The wavefunction in MCTDH is expressed in a two-layer scheme, a first one in terms of time-dependent single-particle basis functions (SPFs, $\{\boldsymbol{\varphi}^{(\kappa)}\}$):

$$\Psi(q_1, \ldots, q_f, t) = \sum_{j_1} \ldots \sum_{j_f} A_{j_1 \cdots j_f}(t) \prod_{\kappa=1}^{f} \varphi_j^{(\kappa)}(q_\kappa, t) \quad (6)$$

and a second in which each SPF is, in turn, expressed in a time-independent basis set ($\{\boldsymbol{\chi}^{(\kappa)}(q_\kappa)\}$):

$$\varphi_{j_\kappa}^{(\kappa)}(q_\kappa, t) = \sum_{i_\kappa=1}^{N_\kappa} c_{j_\kappa i_\kappa}^{(\kappa)}(t) \chi_{i_\kappa}^{(\kappa)}(q_\kappa) \quad (7)$$

the latter, typically, Discrete Variable Representation (DVR) functions (Beck et al., 2000; Light and Carrington, 2000). In this frame, each grid point i_κ (κ-th DOF, $q^{(\kappa)}$) is associated to a localized time-independent basis function ($\chi^{(\kappa)}(q^{(\kappa)})$). Obviously, a minimum number of basis functions, or conversely grid points must exist to achieve the numerical convergence of a given calculation. Such grid representations imply that quantities, particularly the PES, are represented by f-dimensional *tensors*, where f is the number of DOF. If each DOF is represented by 10 grid points, a tensor of 10^f grid points would be necessary to represent the PES. It should be clear at this point that that generation of such a high-dimensional PES tensor directly from electronic structure (i.e., quantum chemistry) codes is, nowadays, a prohibitively-long process.

Apart from diminishing the computational time associated to each quantum chemical calculation, solutions to this issue must imply a reduction in the number of grid points necessary to achieve an accurate grid representation of the PES. This can be achieved in two ways. When considering a more or less localized region of the PES (i.e., centered around a given minimum), local approaches such as the Quartic Force Field representation (QFF) can be used. This is the case when computing vibrational eigenenergies and/or eigenstates (Barone, 2005; Ávila and Carrington, 2009; Neff and Rauhut, 2009). On the other hand, when more global representations are needed (e.g., spectroscopy in multi-well problems, reactivity, etc.) one has to resort to more elaborated forms such as tensor-decomposition algorithms (Kolda and Bader, 2009) or Neural Networks (NN) representations (Manzhos et al., 2006). Two examples of this have been recently proposed for a 6D problem (HONO). With respect to the former, Baranov and Oseledets have used a Tensor-Train tensor-decomposition approach (Baranov and Oseledets, 2015) and Pradhan and Brown have illustrated the use of an exponential NN *ansatz* to represent the same PES (Pradhan and Brown, 2017). In both cases, the number of data-points (i.e., high-level *ab initio* calls) needed to perform the fit was of the order of $\sim 10^4$. Upon an increase of the dimensionality of the problem, this last figure is expected to increase, at least, polynomically, hence preventing the use of these techniques for larger systems.

Our method deals with the aforementioned issues by combining an extremely efficient level of electronic structure, a reparametrized semiempirical Hamiltonian, with an efficient

TABLE 1 | Number and description of the fitting points used in each SRP-optimization stage and the algorithm used in the process.

No. points	Class of points	Type of optimization
53	*core*	Global/Local
367	1D + *core*	Local
546	1D + 2D + *core*	Local
648	1D + 2D + rnd(6D) + *core*	Local
954	1D + 2D + rnd(6D) + LIIC-IRC + *core*	Local
1084	1D + 2D + rnd(6D) + LIIC-IRC + rnd(LIIC) + *core*	Local

*Structures have been generated within a set of fixed boundaries defined in **Table 3**. The initial set of geometries (labeled* core*) consists on 53 points, namely: MIN1, MIN2, TS1, 1D- and 2D-distorted structures using the latter as reference geometries [26 1D, 14 2D], and 10 6D-randomly distorted (*rnd*(6D)) geometries. The number of points at each new set is cumulative. It includes nD-distorted geometries (n = 1, 2, 6), LIIC structures and distortions thereof [noted as* rnd*(LIIC)]. The algorithm for global optimization is MLSL and for local is BOBYQA (see section 2.1). The number of iterations in the global step has been set up to 100,000 and in the local one to 2,000.*

and accurate tensor decomposition scheme, Multigrid POTFIT (MGPF) (Peláez and Meyer, 2013). This tensor decomposition algorithm transforms a multidimensional function (e.g., PES) into Tucker product form (Equation 8) in an *quasi* black-box manner. MGPF, implemented in the MCTDH software package (Worth et al., 2016), avoids running over the full (primitive) MCTDH grid and, instead, uses a series of coarser (nested) grids using a number of PES data-points comparable to the aforementioned methods. However, the big difference is that in our case we shall perform SRP calls, in other words, our *ab initio* method will have the computational cost of a semiempirical one. In fact, as shown by our results (see **Table 1** in section 3.1), we need no more than hundreds of high-level *electronic structure* calls in comparison to the tenths of thousands points required by previous methods. This, obviously, leads to a (small) error inherent to the SRP approximation, but in contrast permits the extension of our approach toward higher-dimensional systems with a little more effort. In the following lines, we shall describe the actual MGPF approach that we have used.

In MGPF, we use a sum-of-products or Tucker expansion for the PES:

$$V = \sum_{j_1,\ldots,j_f}^{[m_1,\ldots,m_f]} C_{j_1,\ldots,j_f} \prod_{\kappa=1}^{f} v_j^{(\kappa)} \tag{8}$$

which, in tensor notation, can be written as: Kolda and Bader (2009)

$$\boldsymbol{V} = \mathcal{C} \times_1 \boldsymbol{v}^{(1)^T} \times_2 \boldsymbol{v}^{(2)^T} \cdots \times_f \boldsymbol{v}^{(f)^T} \tag{9}$$

There $\mathcal{C}$ is the so-called *core* tensor and $\boldsymbol{v}^{(\kappa)}$ are the expansion basis sets for the κ-th DOF. The reader is referred to the original article for a full description of the method and its capabilities (Peláez and Meyer, 2013). More specifically, our current application uses a bottom-up approach to MGPF (Peláez and Meyer, in preparation). The MGPF basis sets ($\{\tilde{\boldsymbol{v}}^{(\kappa)}\}$) can be expressed as:

$$\tilde{\boldsymbol{v}}^{(\kappa)} = \boldsymbol{\rho}^{(\kappa)'} \boldsymbol{\rho}^{(\kappa)^{-1}} \boldsymbol{v}^{(\kappa)} \,. \tag{10}$$

There we have introduced potential density matrices of the form: Peláez and Meyer (2013)

$$\rho_{kk'}^{(\kappa)} := \sum_{I^\kappa} V_{I_k^\kappa} \; V_{I_{k'}^\kappa} \qquad \kappa = 1, \ldots, f \,. \tag{11}$$

where the first index (k) runs along the primitive grid in $\boldsymbol{\rho}^{(\kappa)'}$ and along the coarse one in $\boldsymbol{\rho}^{(\kappa)}$. The transpose of these basis sets reads then:

$$\tilde{\boldsymbol{v}}^{(\kappa)^T} = \boldsymbol{v}^{(\kappa)^T} (\boldsymbol{\rho}^{(\kappa)'} \boldsymbol{\rho}^{(\kappa)-1})^T \tag{12}$$

Substituting in the MGPF expansion V^{MGPF} of the form Equation (9), we unitarily transform both the MGPF basis set ($\tilde{\boldsymbol{v}}$) and the MGPF *core* tensor ($\mathcal{C}$) using the complete basis $\boldsymbol{v}$: Peláez and Meyer (in preparation)

$$\tilde{\boldsymbol{V}}^{\text{MGPF}} = \mathcal{C} \times_1 (\boldsymbol{v}^{(1)^T} \boldsymbol{v}^{(1)}) \tilde{\boldsymbol{v}}^{(1)^T} \times_2 (\boldsymbol{v}^{(2)^T} \boldsymbol{v}^{(2)}) \tilde{\boldsymbol{v}}^{(2)} \cdots \times_f (\boldsymbol{v}^{(f)^T} \boldsymbol{v}^{(f)}) \tilde{\boldsymbol{v}}^{(f)} \tag{13}$$

It should be noted that this transformation does not change the representation. Then one obtains:

$$\tilde{\boldsymbol{V}}^{\text{MGPF}} = \mathcal{V} \times_1 \tilde{\boldsymbol{\gamma}}^{(1)^T} \times_2 \tilde{\boldsymbol{\gamma}}^{(2)^T} \cdots \times_f \tilde{\boldsymbol{\gamma}}^{(f)^T} \tag{14}$$

where $\mathcal{V}$ is the tensor of the energies on the coarse grid and $\tilde{\boldsymbol{\gamma}}^{(\boldsymbol{\kappa})}{=}\boldsymbol{\rho}^{(\boldsymbol{\kappa})'}\boldsymbol{\rho}^{(\boldsymbol{\kappa})\boldsymbol{-1}}$ is the new MGPF basis set. Both quantities, *core* tensor (V) and potential density matrices are directly computed by interfacing the MGPF routine of MCTDH to the openMOPAC software package.

2.4. Calculation of Vibrational Properties: Eigenenergies and Eigenstates

To provide a stringent test to the quality of our series of *chemically accurate* SRP-PES, in addition to RMSEs we have also computed ground and vibrationally excited eigenstates and compared them to those of the reference PES (Richter et al., 2004). These vibrational calculations have been computed using the Heidelberg version of the MCTDH software package (Worth et al., 2016) using our SRP-MGPF PES, as described above. It should be highlighted that the problem we are considering (HONO) features a double well and, consequently, single-reference approaches (e.g., QFF) are not well-suited to its study.

The calculation of the vibrational eigenstates and eigenenergies has been performed by propagating a guess WF in negative imaginary time using the so-called Improved Relaxation method (Meyer and Worth, 2003; Meyer et al., 2006). The MCTDH equations of motion (EOM) are here obtained through a time-independent variational principle. As a result, the propagated configuration interaction coefficients

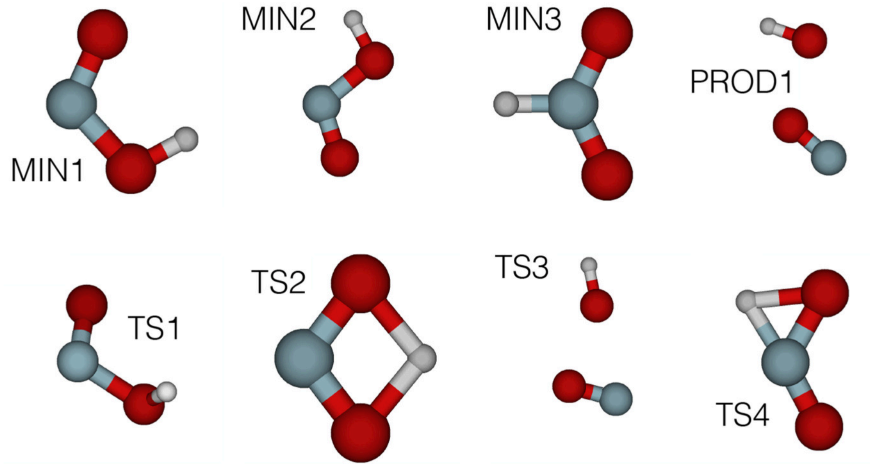


FIGURE 3 | MP2/cc-pVDZ (intermediate HL) structures of HONO automatically obtained using the TSSCDS algorithm on a PM7 (LL) PES. Target geometries in the *cis-trans* isomerization region (MIN1, MIN2, TS1) were subsequently reoptimized at the CCSD(T)/cc-pVQZ (final HL) level of theory.

(A, see Equation 6) are obtained through diagonalization of the Hamiltonian in the basis of the configurations:

$$\sum_L \langle \Phi_J | H | \Phi_L \rangle \, A_L = E \, A_J \,, \tag{15}$$

and the single-particle basis functions (SPFs) are evolved in imaginary time using the *standard* MCTDH EOM (Beck et al., 2000). This iterative procedure is repeated until convergence in the energy. Moreover, a block version of this algorithm, the so-called Block Improved Relaxation, can be used to converge several eigenstates simultaneously, thus leading to the determination of a set of vibrationally excited states.

3. RESULTS AND DISCUSSION

In this section, we present the application of the SRP-MGPF methodology to the actual computation of the HONO (6D) PES for the *cis-trans* isomerization region, which has become a benchmark for this type of studies (Baranov and Oseledets, 2015; Pradhan and Brown, 2017). In the following subsections, we shall discuss the details on the generation of the fitting reference set of points, the reparametrization of the semiempirical Hamiltonian (SRP), and the technical details concerning the direct MGPF tensor decomposition of the SRP-PES into Tucker form. It should be stressed that the novelty and robustness of our approach resides in the fact that requires a minimum intervention of the user, thus qualifying as a *quasi*-black box approach. For the time being, we have interfaced the software openMOPAC to the MCTDH software package through the use of the MGPF tensor decomposition algorithm (Peláez and Meyer, 2013), hence allowing quantum dynamical simulations on a SRP-MGPF PES.

3.1. Computation of the SRP-MGPF PES for the *cis-trans* Isomerization Region in the HONO System (6D)

The first stage in our automated fitting procedure has been the determination of the stationary points of HONO, accomplished through the use of the TSSCDS package (Barnes et al., 2019), as described in section 2.2. Starting from a single random input geometry, LL guess structures have been obtained (see Martínez-Núñez, 2015a,b for a detailed discussion). **Figure 3** presents the corresponding MP2/cc-pvDZ structures. The relevant geometries for our study *cis* (MIN1), *trans* (MIN2) as well as the TS connecting them (TS1) have been reoptimized at the CCSD(T)/cc-pVQZ level of theory. Their geometrical parameters and harmonic frequencies are presented in **Tables S10–S13**. The reason behind the choice of this level of theory is that we have taken as model chemistry the CCSD(T)/cc-pVQZ quality analytical PES of Richter et al. (2004)

The generation of the remaining reference geometries and corresponding energies has been done according to our heuristic approach described in section 2.2. A set of geometries in the form of nD-product grids (n=1, 2) and 6D-random structures have been generated using the three lowest energy stationary points of HONO as pivotal geometries, namely: *cis*, *trans*-conformers and the corresponding TS (see **Figure 3**: MIN1, MIN2, and TS1, respectively). Moreover, the reaction path among them has been taken into account through a piecewise Linear Interpolation in Internal Coordinates (LIIC) (Soto et al., 2006) between the *cis*-TS and TS-*trans* pairs of stationary points (see **Figure S1**) as well as a cloud of distorted structures around them. To ensure that the latter remain close to the *reaction path* (LIIC), each i-th geometry along the LIIC has been generated by distorting along a set of directions resulting from the linear combination of the normal modes of the end structures according to:

$$\Delta\vec{Q}_i = (1 - X_i) \cdot \vec{Q}_{init} + X_i \cdot \vec{Q}_{fin} \qquad \vec{Q} \in \mathbb{R}^{3N-7}$$

where $\vec{Q}_{fin}$ =TS1, $\vec{Q}_{init}$ =MIN1/MIN2. X_i is a number that depends on the *distance* to the end structure. The closer to $\vec{Q}_{fin}$ the more $\Delta\vec{Q}$ resembles the normal modes of the end structure (TS1). Each of our LIIC consists of 50 points and the aforementioned *distance* is simply taken as the ordinal i within the LIIC. It should be noted that the torsion mode has not been included ($3N - 7$ modes in total), since it approximately corresponds to the reaction coordinate. Finally, for a given displacement ($\Delta\vec{Q}$), the geometries around the i-th geometry along the LIIC have been computed as:

$$\vec{R}_i = \vec{R}_i^{(0)} + \sum_{j=1}^{3N-7} f_j \cdot \Delta\vec{Q}_{i,j}$$

where $\vec{R}_i^{(0)}$ is the original geometry of the i-th point of the LIIC, f_j is a small random factor, and $\Delta\vec{Q}_{i,j}$ is the j-th component of $\Delta\vec{Q}_i$.

This systematic manner of generating reference points serves us to control the convergence of the RMSE error at each expansion order, in other words, how insensitive the RMSE is to an increase in the density of points in specific directions (or combinations thereof). This, in turn, provides us with a good estimate of the *lowest possible* number of reference geometries at each stage. In **Table 1**, we present the different convergence stages in terms of number of fitting points used together with the associated optimization algorithm. As it can be observed, at each specific stage, we either increase the density of points in the indicated directions (*modes/coordinates*) or add a new class of points in the form of a LIIC, for instance.

The first stage consists on a global optimization (MLSL) followed by a local one (BOBYQA) using a small number of judiciously chosen points: the RXN and a cloud of random geometries around them, adding up to a total of 53 points. This has enabled a very large number of iterations (10^5). The underlying hypothesis behind this calculation is that a reasonable and cheap estimate of the *global minimum* (set of SRP parameters yielding the minimum RMSE) can be obtained. Our best set of parameters at this stage ($\boldsymbol{\zeta}_{53}$, where 53 is the number of fitting points) yielded an initial RMSE of 806.8 cm^{-1} (**Table S1**). In the subsequent stages, we have performed local optimizations (BOBYQA) with 2,000 iterations. Before proceeding any further, we would like to justify the use of a global algorithm exclusively at the first stage, in other words, ζ_{53} must indeed correspond to a set near the *global* minimum or a local deep minimum. First, from a computational perspective, it should be noted that a small number of fitting points is ideally suited for this task. Second, we have performed calculations justifying this fact. **Table S2** (column 2) presents the BOBYQA variation of the RMSE for an increasing number of 1D-sets of fitting points. It can be observed that upon increase of this number, from 192 until 2088 fitting points, the RMSE monotonically decreases from 482.13 cm^{-1} till 365.13 cm^{-1}. According to our reasonings above, one should take the SRP parameters of the last set of points ($\boldsymbol{\zeta}_{1542}$ or $\boldsymbol{\zeta}_{2088}$) corresponding to the best RMSE of the 1D-series. For the sake of efficiency, we considered the $\boldsymbol{\zeta}_{1542}$. With this set of SRP parameters, we recomputed the whole series of RMSEs for the different sets of 1D-points and we observed a very close agreement with the BOBYQA values, except for the 192 set. This shows that indeed all sets of parameters of this series (from $\boldsymbol{\zeta}_{367}$ on) lie within the *same* RMSE landscape region (see **Figure 4**) and, in turn, validates our initial approach with a small number of *representative* points. One can then safely conclude that just 367 fitting points are necessary to improve the SRP-fitting at the 1D-level. Hence, subsequent 2D optimizations will start with the ($\boldsymbol{\zeta}_{367}$) set. A detailed description of all stages and RMSE values is presented in **Tables S1–S9**. A somewhat more complete information can be obtained through the cumulative error computed by addition of the RMSEs resulting form the configurations up to a certain energy value (see **Figure 5**). It can observed that for all sets of parameters, with the exception of $\boldsymbol{\zeta}_{53}$, the RMSEs remain below the limit of chemical accuracy (1 kcal/mol≈ 350 cm^{-1}) within the targeted PES region (*cis-trans* isomerization). Moreover, in the last stage we have removed all structures with energies above 5000 cm^{-1} (above the classical barrier) and included an extra set of random points around the stationary points. This new set of points has been used to BOBYQA reoptimize the SRP. We observe a clear improvement of the RMSE in such a way that, up to 8000 cm^{-1}, the RMSE is inferior to the chemical accuracy level. The correctness of these results has been supported by a calculation using a validation set consisting of 1200 6D random points with energies below 12000 cm^{-1} for which the same pattern is obtained. We have also compared the geometries and harmonic frequencies of all stationary points at the reference *ab initio* level of theory and at the SRP level for each stage. Geometries are displayed in **Tables S10–S12** and harmonic frequencies are shown in **Table 2**. As it can be observed, SRP does indeed improve, in terms of both geometrical parameters and harmonic frequencies, with respect to the original PM7 and, furthermore, we obtain a very good agreement with the reference *ab initio* data. This is particularly true for the last stage ($\boldsymbol{\zeta}_{1084}$).

To finalize this section, we present in **Figure 6** a comparison of 2D projections of the *cis-trans* isomerization regions for: (i) the reference surface, (ii) the SRP-PES($\boldsymbol{\zeta}_{1084}$); and (iii) the PM7 semiempirical Hamiltonian. These contour plots have been obtained through orthogonalization of the two LIIC vectors used in **Figure 6**. The positive effect of the reparametrization can be clearly observed: while PM7 provides a *blurred* description of the TS region, the SRP-PES reproduces it correctly.

3.1.1. Classical Molecular Dynamics on the SRP-PES

As a first test of the quality of the SRP-PES, we have carried out classical molecular dynamics simulations for the HONO system in full dimensionality using the VENUS96 software package (Hu et al., 1991). Classical trajectories have been run using the reference PES (Richter et al., 2004). The energies of the so-obtained geometries have been subsequently computed at the SRP-PES level and compared to the original calculation. Starting from the equilibrium geometries of the *cis* and *trans* isomers, we have propagated for 1 *ps* each trajectory with a time-step of 5*fs*. The vibrational energy of each starting geometry was classically distributed in a random way between all normal modes using the option *normal mode sampling* of the VENUS

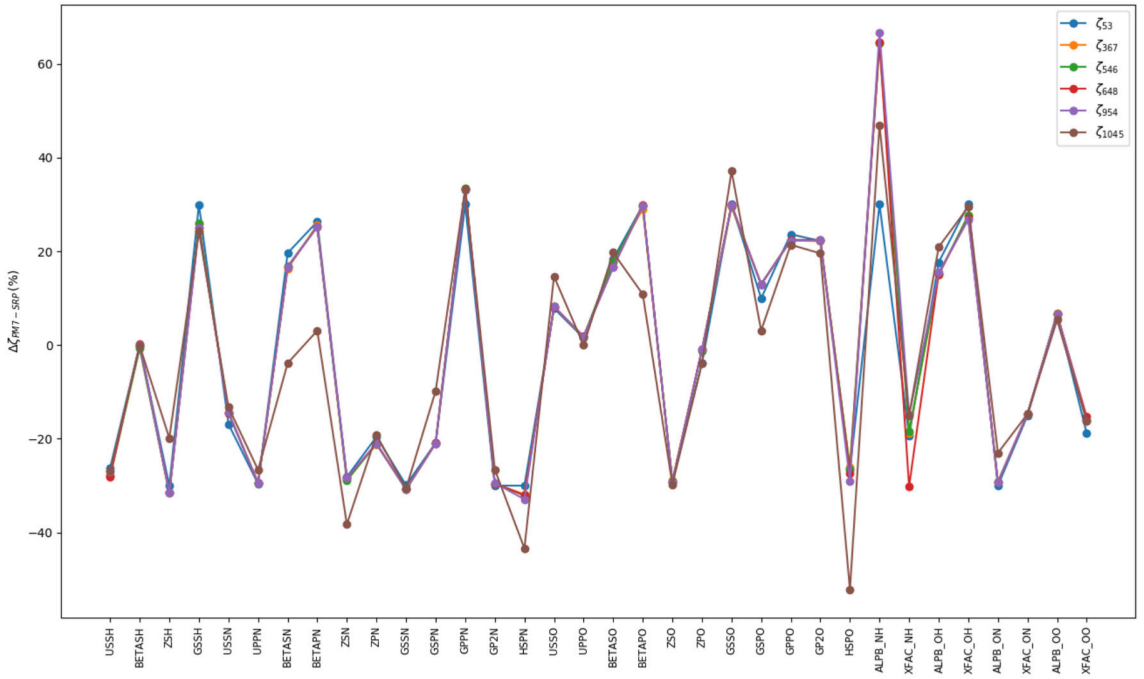


FIGURE 4 | Percentage of variation of the SRP parameters with respect to the original PM7 ones. Each fitting stage is represented by its *optimal* parameters, ζ_N, where N is the number of points used in the process (see **Table 1**). On abscissas we present the label of semiempirical parameters for the different type of atoms in HONO. Standard semiempirical parameter labeling has been used (Stewart, 2013). Parameters from USSH until HSPO correspond to a single type of atom whereas parameters labeled ALPB_{XY} and XFAC_{XY} correspond to two-atom ones (atom X and atom Y).

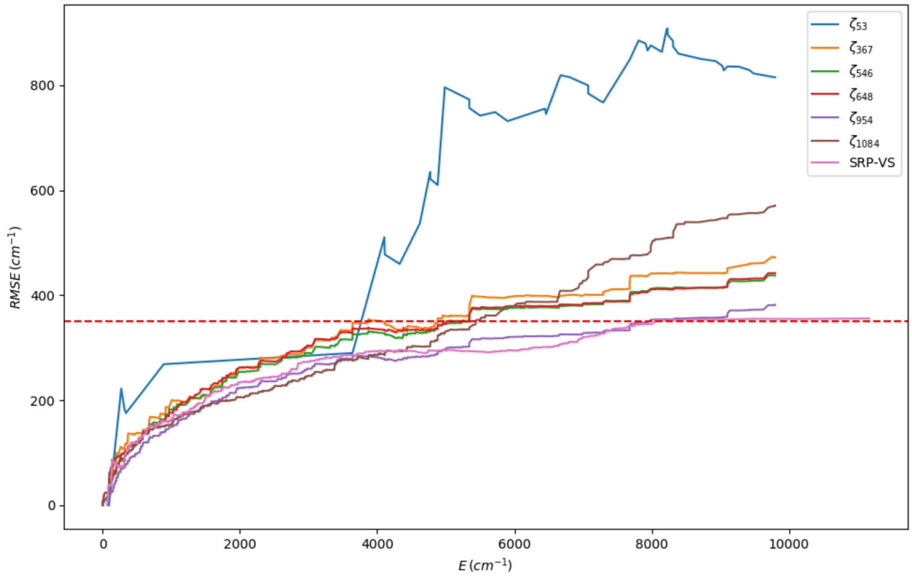


FIGURE 5 | Cumulative RMSE for each SRP-fit labeled by its set of parameters, ζ_N, where N is the number of points used in the fit (see **Table 1**). The last set (ζ_{VS}) corresponds to the validation set. The red dotted horizontal line represents the value of the chemical accuracy (1 kcal/mol≈ 350 cm^{-1}).

software. We have computed 10 trajectories per isomer, each isomer having 4 different vibrational energies (5, 10, 15, and 20 kcal/mol) thus making a total of 80 trajectories and 16,080 geometries. In **Figure 7**, we present a comparison of the variation of the potential energies along two of these trajectories. As it can be observed, the PM7 largely deviates from the reference calculation both in their relative values and the phase, whereas SRP-PES follows closely the *ab initio* values. In particular, it

TABLE 2 | Harmonic frequencies of the normal modes of each stationary point at the CCSD(T)/cc-pVQZ *ab initio* level of theory and corresponding values for the PM7 method and the SRPs in the different stages of the optimization.

	Harmonic frequencies (cm^{-1})							
	Ab initio	ζ_{PM7}	ζ_{53}	ζ_{367}	ζ_{546}	ζ_{648}	ζ_{954}	ζ_{1084}
TS	-599.2	-553.6	-581.0	-565.1	-568.7	-570.3	-573.3	-606.8
	559.1	621.7	512.2	467.1	465.5	463.6	463.8	597.2
	791.2	1021.3	654.7	649.5	652.7	653.6	654.9	738.4
	1122.3	1175.3	1174.3	1092.3	1099.8	1095.8	1106.8	1195.4
	1728.0	1839.5	1763.8	1705.8	1709.4	1710.5	1711.0	1737.9
	3785.3	2801.7	3747.9	3568.9	3585.3	3585.0	3586.4	3736.2
cis	648.7	589.0	615.4	613.0	616.1	618.2	619.3	622.5
	687.9	629.2	724.9	712.0	718.4	716.0	718.4	698.9
	901.9	1084.8	745.3	715.4	721.0	721.7	728.4	854.2
	1350.9	1346.0	1316.4	1252.5	1255.9	1253.7	1262.3	1369.5
	1675.5	1823.5	1725.5	1693.2	1696.6	1698.3	1701.6	1719.1
	3632.1	2802.9	3668.9	3504.6	3519.8	3520.0	3521.4	3667.3
trans	574.8	455.9	517.1	515.2	517.1	518.1	521.6	540.5
	633.1	609.8	533.3	515.2	519.1	523.7	528.3	602.5
	839.6	1096.0	730.5	736.6	741.7	744.7	748.9	835.1
	1319.3	1308.8	1232.9	1130.0	1136.9	1131.6	1148.4	1264.6
	1732.6	1826.5	1715.9	1666.7	1670.2	1671.9	1674.8	1704.7
	3790.8	2828.3	3815.8	3662.7	3678.9	3680.9	3682.9	3796.1

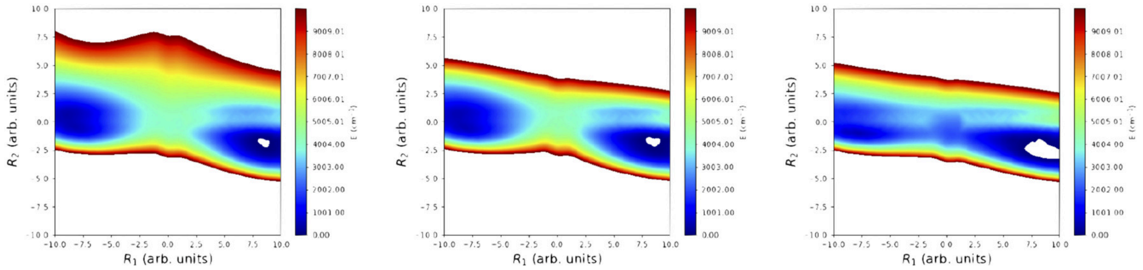


FIGURE 6 | Comparison of the 2D projections of the cis-trans isomerization region for: (i) reference PES (Richter et al., 2004) (left panel); (ii) SRP-PES (ζ_{1084}) (middle panel); and (iii) PM7-PES (right panel). These projections have been obtained by orthonormalization of two linear interpolation (LIIC) vectors as described in Soto et al. (2006).

is remarkable the fact that for low energies PM7 presents a large amount of structures with energies below the value of the global minimum, the *trans* conformer. To finalize this subsection, we would like to provide some performance features of the SRP-PES which directly show the efficiency of the underlying openMOPAC software. In the case of the HONO, from an average of the order of $\sim 10^4$ points, we have obtained a mean CPU-time of 10^{-2} s per single-point energy. Moreover, Hessians are computed in less than a second. This properties make SRP approaches suitable for any on-the-fly type of calculation. In particular, we are currently exploring their use with non-grid based quantum dynamical methods such as the Direct-Dynamics Variational Multiconfigurational Gaussian (DD-vMCG) method (Richings et al., 2015).

3.2. Full Quantum Analysis of the Vibrational Properties of the SRP-PES for the *cis-trans* HONO System (6D)

To further assess the quality of our SRP-PES we have computed vibrational properties by means of MCTDH quantum dynamical calculations and the results have been compared to the ones from the reference PES (Richter et al., 2004). More specifically, ground and excited vibrational states as well as vibrational spectra, in the form of Fourier transforms of autocorrelation functions. At this point, it should be recalled that our main goal is not to achieve spectroscopical accuracy but to provide PESs, in a fully automated fashion, accurate enough to disentangle chemical processes.

3.2.1. MGPF Tensor Decomposition of the HONO 6D PES

To *interface* the SRP-PES with the MCTDH quantum dynamics software package, we have used the Multigrid POTFIT tensor decomposition algorithm (Peláez and Meyer, 2013). More specifically, all PES *calls* within the MGPF workflow have been addressed directly to the openMOPAC software package using an external set of *optimal* SRP parameters. In other words, at each grid point, i.e., configuration, a SCF process is performed. Of course, this is only possible due to the high efficiency of the underlying PM7 frame. This fact, precisely, has allowed us to circumvent the issues encountered in previous studies in which the *ab initio* energies were generated directly from a quantum chemical calculation thus severely limiting the level of theory which could be applied.

We have carried out *bottom-up* MGPF calculations Peláez and Meyer (2013) to the different SRP-PESs at different parameter optimization stages. In **Table S14**, we present a comparison in terms of CPU time and memory needs for a reference exact Tucker decomposition (using POTFIT, PF) (Jäckle and Meyer, 1996) and the different MGPF tensor decomposition levels that we have used in this work. The full primitive grid, needed in PF, consists of $2.804 \cdot 10^7$ points. In contrast, the coarse grids in MGPFs include every third, fourth, or fifth fine grid point for each DOF. These coarse grids have been labeled *ev*3, *ev*4, and *ev*5 and consist of 172,800, 51,200, and 18,432 coarse grid points, respectively. The MGPF partial grids increase these figures by a factor $<$10. This is due to the fact that the contracted mode lies fully in the fine grid (see section IIIB in Peláez and Meyer, 2013). Hence, as expected, MGPF is orders of magnitude less demanding that an exact decomposition. The global RMSE values show that MGPF PES are accurate, cheap and, more importantly, add a very small (global, full grid) error to the PES. Finally, it should also be highlighted that none of our SRP-PES present energies below the global minimum (*trans* conformer), whereas the PM7 does. In other words, PM7 presents artificial PES structure when compared to the reference one. We have observed that even the simplest SRP optimization corrects this wrong behavior.

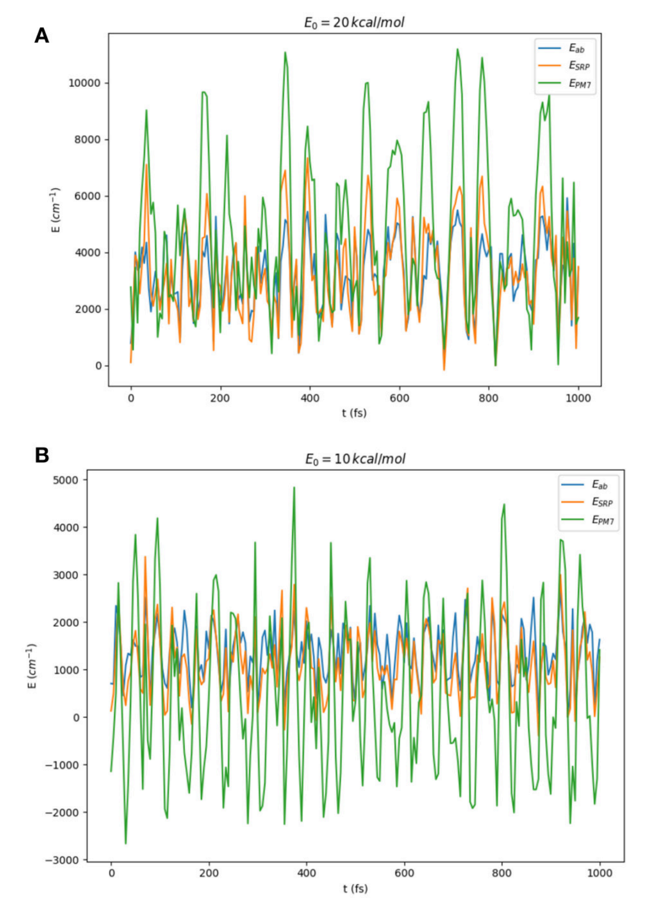


FIGURE 7 | Comparison of ab initio (blue line), PM7 (green line), and SRP-PES (ζ_{1084}) (orange line) energies for the geometries generated in classical on-the-fly trajectories of HONO(6D) with total energies (randomly distributed among all modes) of 10 and 20 kcal/mol starting at: **(A)** the *trans*-conformer and **(B)** the *cis*-conformer.

3.2.2. MCTDH Quantum Molecular Dynamics on the SRP-MGPF

As discussed in section 2.3, MCTDH requires the discretisation of the configuration space. The HONO (6D) molecule has been represented in internal coordinates (see **Figure 8**) as in previous works (Peláez and Meyer, 2013; Pradhan and Brown, 2017), and a Discrete Variable Representation (DVR) grid has been defined accordingly (see **Table 3**). We have performed ground and excited eigenstate vibrational calculations for the reference PES, the PM7-MGPF PES as well as for selected SRP-MGPF PES using the Improved Relaxation algorithm and its Block version, as implemented in the Heidelberg version of MCTDH (Meyer et al., 2006). We have combined the physical modes into logical particles as follows: [ϕ=15], [d_{OH}=10] [u_2, d_{ON}=25], [u_1, d_{NO}=25], where the number represents the number of single-particle functions (SPFs) and $u_i = \cos\theta_i$ (see **Figure 8**). In all cases, the initial wave packet has been propagated in *negative imaginary time* (see section 2.4) during 500 fs.

With respect to ground state energies, the reference PES yields a value of 4367.7 cm^{-1} for the Zero Point Energy (ZPE) and the PM7-MGPF PES a value of 3221.3 cm^{-1}, well off the analytical one. We attribute this discrepancy to the artificial structure of

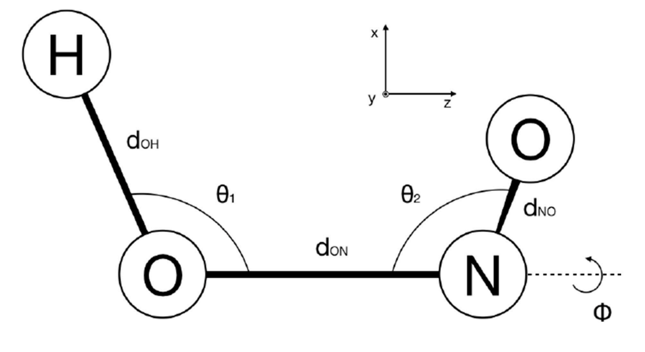


FIGURE 8 | Definition of the internal coordinates of HONO used in this work.

the PES revealed by the presence of *negative* energies (geometries with energies below the global minimum, *trans* conformer) as discussed in section 3.2.1) and clearly illustrated in **Figure 7**. On the other hand, concerning the SRP-MGPF PESs, a nice convergence can be observed upon increase of the number of fitting points, toward a final value of 4332.8 cm^{-1} which compares well with the analytical one. It is also remarkable that a simple fit using only 53 fitting points already leads to a qualitative improvement with respect to PM7. Moreover, our results show that the ZPE values are somewhat insensitive to the size of the coarse grid (cf. last three rows of **Table 4**). Consequently, we shall use hereafter the *ev*5 SRP-MGPF scheme.

We have also computed the 20 lowest-lying vibrational eigenstates of HONO (**Table 5**). It should be noted that this energy interval spans all HONO fundamentals except the OH stretching mode. For this, we have considered four different PES, namely: (i) PM7-MGPF, SRP-MGPF with $\boldsymbol{\zeta}_{53}$ and $\boldsymbol{\zeta}_{1084}$, as well as the reference (exact) PES. The first remark to be done is that the original PM7-MGPF PES fails to predict the initial vibrational state corresponding to the *ground state* of the *cis* conformer (Richter et al., 2004). In contrast, even at the minimum level of reparametrization ($\boldsymbol{\zeta}_{53}$), this eigenstate is obtained. Furthermore, this incorrect behavior worsens upon increase of the energy. In fact, eigenenergies are off by several hundreds of cm^{-1} in almost the its whole range. This can be readily understood by simple observation of the 2D contour plots of the *cis-trans* region of the PES (see **Figure 6**). In contrast, both SRP-MGPFs nicely follow the reference values and, what is more important, the discrepancies (of the order of tens of cm^{-1}) do not increase but remain, in average, constant.

Finally, to take into account higher excited vibrational states, we have computed a vibrational spectrum by Fourier transform

TABLE 3 | Definition of the MCTDH primitive grid: HO denotes a harmonic oscillator (Hermite) and cos a cosine Discrete Variable Representation (DVR) basis functions.

DOF	DVR	N	Range
d_{OH}	HO	18	[1.30, 2.45]
d_{NO}	HO	13	[1.90, 2.60]
u_2	HO	13	[-0.65, -0.10]
d_{ON}	HO	16	[2.10, 3.25]
u_1	HO	18	[-0.65, 0.25]
ϕ	cos	32	[0, $2\pi/2$]

*N is the number of primitive (fine) grid points. The range represents the first and last grid points in atomic units for the distances and ϕ is the torsion angle in radians. Cosines of the valence angles have been used: $u_i = \cos\theta_i$. See **Figure 8** for the definition. Physical degrees of freedom have been combined into logical modes or* particles *according to the following scheme: [ϕ], [d_{OH}] [u_2, d_{ON}], [u_1, d_{NO}]. The first particle (ϕ) has been contracted in MGPF (see section IIIB in Peláez and Meyer, 2013).*

TABLE 4 | Ground state energies of HONO using PESs of different quality.

Set	MGPF	ZPE (cm^{-1})
$\boldsymbol{\zeta}_{PM7}$	*ev*4	3221.3
$\boldsymbol{\zeta}_{53}$	*ev*4	4070.7
$\boldsymbol{\zeta}_{648}$	*ev*4	4095.0
$\boldsymbol{\zeta}_{1084}$	*ev*4	4332.8
$\boldsymbol{\zeta}_{1084}$	*ev*5	4330.8
$\boldsymbol{\zeta}_{1084}$	*ev*3	4332.9

*The first column indicates the set of SRP parameters used, labeled by its set of parameters, ζ_N, where N is the number of points used in the fit (see **Table 1**). The second column presents the size of the MGPF coarse grid: evn indicates a coarse grid in which every (ev) n-th fine grid point has been considered (see section 3.2.1). The final column presents the Zero Point Energies (ZPE) for each of the previous PES.*

TABLE 5 | Comparison of the 20 lowest vibrational eigenvalues of HONO for different PESs denoted by its set of parameters, $\boldsymbol{\zeta}_N$, where N is the number of points used in the fit (see **Table 1**).

	Vibrational eigenenergies (cm^{-1})			
	$\boldsymbol{\zeta}_{PM7}$	$\boldsymbol{\zeta}_{53}$	$\boldsymbol{\zeta}_{1084}$	**Analytical**
	0.0	0.0	0.0	0.0
	593.6	163.0	88.5	94.1
	794.3	604.7	597.1	600.8
	1070.6	693.2	703.9	710.7
	1151.5	706.9	822.3	795.9
	1186.3	888.9	917.9	944.1
	1365.9	1134.3	1012.5	1055.4
	1403.1	1204.8	1189.7	1188.1
	1641.3	1221.6	1234.7	1264.9
	1659.6	1263.0	1317.9	1306.6
	1751.1	1308.9	1363.5	1312.8
	1773.1	1361.6	1417.2	1385.3
	1811.5	1395.7	1451.1	1404.8
	1869.9	1424.9	1530.5	1547.9
	1968.7	1426.3	1607.7	1574.9
	2011.4	1612.4	1633.9	1640.9
	2060.3	1656.9	1690.9	1689.9
	2118.1	1698.3	1743.0	1726.0
	2136.5	1748.6	1778.7	1762.4
	2226.5	1842.0	1785.8	1779.7
	2253.3	1853.0	1807.3	1829.0
RMSE	360.2	58.4	24.5	–
	N/A	–	[42.0]	–
MAD	53.7	38.3	23.7	–
	N/A	–	[25.5]	–

Energies have been computed by MCTDH Block Improved Relaxation (see section 2.4). All PESs have been MGPFitted *using a coarse grid consisting on 18,432 points, the so-called ev5 (see section 3.2.1). The first column presents the PM7-MGPF values (PM7), second and third correspond to SRP-MGPF with $\boldsymbol{\zeta}_{53}$ and $\boldsymbol{\zeta}_{1084}$, respectively. The last column presents the corresponding eigenenergies obtained using the analytical surface by Richter et al. (2004). The last four rows present the RMSE and the mean-absolute deviation (MAD) of each set of eigenvalues with respect to the analytical ones. The values in square brackets indicate the RMSE and MAD values taking into account the corresponding OH stretching anharmonic frequencies. The latter have been obtained through Fourier transform of an autocorrelation function (see **Figure 9**): (i) Analytical: 3533.8 cm^{-1} and (ii) $\boldsymbol{\zeta}_{1084}$: 3695.7 cm^{-1}. It should be noted that the PM7 values could not be determined (indicated by N/A) owing to a wrong behavior of the PM7-PES at this energy range (see **Figure 9**).*

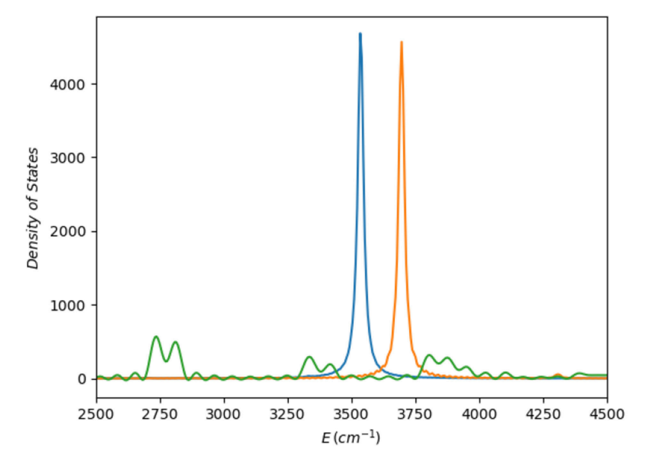


FIGURE 9 | Vibrational spectra computed as the Fourier transform of the autocorrelation function obtained after excitation of one quantum in the OH stretching vibration centered the *cis* conformer region: (i) green line corresponds to the PM7-MGPF PES; (ii) orange line to the SRP-MGPF (ζ_{1084}) PES; and (iii) blue line, the reference PES (*ab initio*) (Richter et al., 2004).

of the autocorrelation function corresponding to the dynamics of a wave packet generated by excitation of a quantum of energy in the OH stretching mode in the *cis* region of the potential. As observed (**Figure 9**), the PM7-MGPF spectrum is radically different to that of the reference PES, whereas the SRP-MGPF one shows the correct behavior. Apart from the, certainly not unexpected, shift in energy, both reference PES and SRP-MGPF reveal that the OH mode is practically uncoupled from the rest.

4. CONCLUSIONS AND FUTURE PROSPECTS

We have introduced Specific Reaction Parameter Multigrid POTFIT (SRP-MGPF) a methodology which permits the generation of global chemically accurate Potential Energy Surfaces in sums-of-products (Tucker) form in a *quasi* black-box manner starting from a random input geometry. The SRP-MGPF workflow combines: (i) the automated determination of stationary points of a Potential Energy Surface (PES); (ii) the reparametrization of a Semiempirical Hamiltonian (SRP) using high-level *ab initio* data; and (iii) direct tensor-decomposition of the resulting SRP-PES with the Multigrid POTFIT (MGPF) algorithm. The resulting surface can be used with any on-the-fly dynamical software or, after MGPF, with grid-based quantum dynamical method, in particular the Multiconfiguration Time-Dependent Hartree (MCTDH) method. We have proven the validity of this method by fitting the SRP-MGPF PES for the HONO system in full dimensionality (6D) and reproducing, to a good agreement, the vibrational properties of a surface of CCSD(T)/cc-pVQZ quality. Current work deals with the extension of the method to treat coupled electronic excited states. To finalize, it should be highlighted that SRP-MGPF provides an inexpensive and accurate enough means of performing full-dimensional chemically meaningful quantum or classical simulations.

DATA AVAILABILITY

All datasets generated for this study are included in the manuscript/**Supplementary Files**.

AUTHOR CONTRIBUTIONS

DP conceived the original idea, managed the project, and wrote the first draft of the manuscript. RP-B and DP were responsible for its implementation. RP-B was main responsible for the development of the necessary software. EM-N has been responsible for the TSSCDS calculations and classical trajectories. All authors contributed to manuscript revision, read and approved the submitted version.

FUNDING

RP-B thanks the Région Hauts-de-France and the Université de Lille for a Ph.D. Fellowship. RP-B and DP thank the Région Hauts-de-France and the Ministére de l'Enseignement Supérieur et de la Recherche (CPER Climibio), and the European Fund for Regional Economic Development for their financial support. This work has been partially supported by the Agence Nationale de la Recherche through the Labex CaPPA (ANR-11-LABX-0005-01). EM-N thanks Xunta de Galicia (Research Grant No. ED431C 2017/17). EM-N is very grateful to l'Institut de Physique for a CNRS visiting professorship to the Laboratoire PhLAM of the University of Lille.

ACKNOWLEDGMENTS

The authors thank Dr. James Stewart for helpful and lively discussions as well as for having granted us permission for using the openMOPAC package as default in SRP-MGPF. DP thanks Prof. Dr. Diego Troya for discussions on reparametrizations of semiempricial Hamiltonians. DP would like to thank Prof. Dr. Hans-Dieter Meyer for encouraging discussions. The authors thank Dr. Alejandro Rivero Santamaría for discussions on the NLOpt libraries and Prof. Dr. Maurice Monnerville for fruitful scientific exchanges. The authors thank Dr. Sabine Kopec for providing the initial LL-TSSCDS calculations and Hugo Geindre for the use of his script for the generation of random geometries based on normal modes.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2019.00576/full#supplementary-material

REFERENCES

Ackley, D. H. (1987). *A Connectionist Machine for Genetic Hillclimbing*. Norwell, MA: Kluwer Academic Publishers.

Ávila, G., and Carrington, T. Jr. (2009). Nonproduct quadrature grids for solving the vibrational Schrödinger equation. *J. Chem. Phys.* 131:174103. doi: 10.1063/1.3246593

Baranov, V., and Oseledets, I. (2015). Fitting high-dimensional potential energy surface using active subspace and tensor train (AS+TT) method. *J. Chem. Phys.* 143:174107. doi: 10.1063/1.4935017

Barnes, G. L., Kopec, S., Peláez, D., Rodríguez, A., Rodríguez-Fernández, R., J. J. P., Stewart, P. T., et al. (2019). Available online at: https://rxnkin.usc.es/index.php/AutoMeKin

Barone, V. (2005). Anharmonic vibrational properties by a fully automated second-order perturbative aproach. *J. Chem. Phys.* 122:014108. doi: 10.1063/1.1824881

Beck, M. H., Jäckle, A., Worth, G. A., and Meyer, H.-D. (2000). The multi-configuration time-dependent Hartree (MCTDH) method: a highly efficient algorithm for propagating wave packets. *Phys. Rep.* 324, 1–105. doi: 10.1016/S0370-1573(99)00047-2

Braams, B. J., and Bowman, J. M. (2009). Permutationally invariant potential energy surfaces in high dimensionality. *Int. Rev. Phys. Chem.* 28, 577–606. doi: 10.1080/01442350903234923

Burghardt, I., Giri, K., and Worth, G. A. (2008). Multi-mode quantum dynamics: the G-MCTDH method applied to the absorption spectrum of pyrazine. *J. Chem. Phys.* 129:174104. doi: 10.1063/1.2996349

Christensen, A. S., Kubar, T., Cui, Q., and Elstner, M. (2016). Semiempirical quantum mechanical methods for noncovalent interactions for chemical and biochemical applications. *Chem. Res.* 116, 5301–5337. doi: 10.1021/acs.chemrev.5b00584

Christiansen, O. (2004). Vibrational coupled cluster theory. *J. Chem. Phys.* 120, 2149–2159. doi: 10.1063/1.1637579

Dawes, R., Thompson, D. L., Guo, Y., Wagner, A. F., and Minkoff, M. (2007). Interpolating moving least-squares methods for fitting potential energy surfaces: computing high-density potential energy surface data from low-density ab initio data points. *J. Chem. Phys.* 126:184108. doi: 10.1063/1.2730798

Frankcombe, T. J., and Collins, M. A. (2011). Potential energy surfaces for gas-surface reactions. *Phys. Chem. Chem. Phys.* 13, 8379–8391. doi: 10.1039/c0cp01843k

Gatti, F. (ed.). (2014). Molecular Quantum Dynamics. Heidelberg: Springer.

Haller, A., Peláez, D., and Bande, A. (2019). Inter-coulombic decay in laterally-arranged quantum dots controlled by polarized lasers. *J. Phys. Chem. C* 123, 14754–14765. doi: 10.1021/acs.jpcc.9b01250

Hu, X., Hase, W. L., and Pirraglia, Y. (1991). Vectorization of the general monte carlo classical trajectory program venus. *J. Comput. Chem.* 12, 1014–1024. doi: 10.1002/jcc.540120814

Jäckle, A., and Meyer, H.-D. (1996). Product representation of potential energy surfaces. *J. Chem. Phys.* 109:3772. doi: 10.1063/1.471513

Jiang, B., and Guo, H. (2014). Permutation invariant polynomial neural network approach to fitting potential energy surfaces. III. Molecule-surface interactions. *J. Chem. Phys.* 141:034109. doi: 10.1063/1.4887363

Johnson, S. G. (2011). *The NLopt Non-linear-optimization Package*. Available online at: http://ab-initio.mit.edu/nlopt

Jones, E., Oliphant, T., and Peterson, P. (2001). *SciPy: Open Source Scientific Tools for Python*. Available online at: http://www.scipy.org/

Kan, A. H. G., and Timmer, G. T. (1987). Stochastic global optimization methods part II: multi level methods. *Math. Program.* 39, 57–78. doi: 10.1007/BF02592071

Kolda, T. G., and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Rev.* 51, 455–500. doi: 10.1137/07070111X

Kopec, S., Martínez-Núñez, E., Soto, J., and Peláez, D. (2019). vdW-TSSCDS - an automated and global procedure for the computation of stationary points on intermolecular potential energy surfaces. *Int. J. Quant. Chem.* 2019:e26008. doi: 10.1002/qua.26008

Kosloff, R. (1988). Time-dependent quantum-mechanical methods for molecular dynamics. *J. Phys. Chem.* 92:2087. doi: 10.1021/j100319a003

Layfield, J. P., Owens, M. D., and Troya, D. (2008). Theoretical study of the dynamics of the H+CH_4 and H+C_2H_6 reactions using a specific-reaction-parameter semiempirical hamiltonian. *J. Phys. Chem.* 128:194302. doi: 10.1063/1.2918358

Levine, B. G., Coe, J. D., Virshup, A. M., and Martínez, T. J. (2008). Implementation of *ab initio* multiple spawning in the Molpro quantum chemistry package. *Chem. Phys.* 347, 3–16. doi: 10.1016/j.chemphys.2008.01.014

Light, J. C., and Carrington, T. Jr. (2000). Discrete variable representations and their utilization. *Adv. Chem. Phys.* 114, 263–310. doi: 10.1002/9780470141731.ch4

Manzhos, S., and Carrington, T. Jr. (2006). Using neural networks to represent potential surfaces as sum of products. *J. Chem. Phys.* 125:084109. doi: 10.1063/1.2387950

Manzhos, S., Wang, X., Dawes, R., and Carrington, T. Jr. (2006). A nested molecule-independent neural network approach for high-quality potential fits. *J. Phys. Chem. A* 110, 5295–5304. doi: 10.1021/jp055253z

Martínez-Núñez, E. (2015a). An automated method to find transition states using chemical dynamics simulations. *J. Comp. Chem.* 36:222. doi: 10.1002/jcc.23790

Martínez-Núñez, E. (2015b). An automated transition state search using classical trajectories initialized at multiple minima. *Phys. Chem. Chem. Phys.* 17:14912. doi: 10.1039/C5CP02175H

Meyer, H.-D., Le Quéré, F., Léonard, C., and Gatti, F. (2006). Calculation and selective population of vibrational levels with the Multiconfiguration Time-Dependent Hartree (MCTDH) algorithm. *Chem. Phys.* 329, 179–192. doi: 10.1016/j.chemphys.2006.06.002

Meyer, H.-D., and Worth, G. A. (2003). Quantum molecular dynamics: propagating wavepackets and density operators using the multiconfiguration time-dependent Hartree (MCTDH) method. *Theor. Chem. Acc.* 109, 251–267. doi: 10.1007/s00214-003-0439-1

Neff, M., and Rauhut, G. (2009). Toward large scale vibrational configuration interaction calculations. *J. Chem. Phys.* 131:124129. doi: 10.1063/1.3243862

Peláez, D., and Meyer, H.-D. (2013). The multigrid POTFIT (MGPF) method: grid representations of potentials for quantum dynamics of large systems. *J. Chem. Phys.* 138:014108. doi: 10.1063/1.4773021

Peláez, D., and Meyer, H.-D. (2017). On the infrared absorption spectrum of the hydrated hydroxide ($H_3O_2^-$) cluster anion. *Chem. Phys.* 482:100. doi: 10.1016/j.chemphys.2016.08.025

Peláez, D., Sadri, K., and Meyer, H.-D. (2014). Full-dimensional MCTDH/MGPF study of the ground and lowest lying vibrational states of the bihydroxide $H_3O_2^-$ complex. *Spectrochim. Acta A* 119, 42–51. doi: 10.1016/j.saa.2013.05.008

Powell, M. J. D. (2009). *The BOBYQA Algorithm for Bound Constrained Optimization without Derivatives*. Cambridge NA Report NA2009/06. Cambridge: University of Cambridge, 26–46.

Pradhan, E., and Brown, A. (2017). A ground state potential energy surface for HONO based on a neural network with exponential fitting functions. *Phys. Chem. Chem. Phys.* 19:22272. doi: 10.1039/C7CP04010E

Puzzarini, C., Bloino, J., Tasinato, N., and Barone, V. (2019). Accuracy and interpretability: the devil and the holy grail. new routes across old boundaries in computational spectroscopy. *Chem. Res.* 119, 8131–8191. doi: 10.1021/acs.chemrev.9b00007

Rauhut, G. (2004). Efficient calculation of potential energy surfaces for the generation of vibrational wave functions. *J. Chem. Phys.* 121:9313. doi: 10.1063/1.1804174

Rauhut, G. (2007). Configuration selection as a route towards efficient vibrational configuration interaction calculations. *J. Chem. Phys.* 127:184109. doi: 10.1063/1.2790016

Richings, G. W., and Habershon, S. (2018). MCTDH on-the-fly: efficient grid-based quantum dynamics without pre-computedpotential energy surfaces. *J. Chem. Phys.* 148:134116. doi: 10.1063/1.5024869

Richings, G. W., Polyak, I., Spinlove, K. E., Worth, G. A., Burghardt, I., and Lasorne, B. (2015). Quantum dynamics simulations using gaussian wavepackets: the vmcg method. *Int. Rev. Phys. Chem.* 34:269. doi: 10.1080/0144235X.2015.1051354

Richter, F., Hochlaf, M., Rosmus, P., Gatti, F., and Meyer, H.-D. (2004). A study of mode–selective trans–cis isomerisation in HONO using ab initio methodology. *J. Chem. Phys.* 120, 1306–1317. doi: 10.1063/1.1632471

Rios, L. M, and Sahinidis, N. V. (2013). Derivative-free optimization: a review of algorithms and comparison of software implementations. *J. Global Optim.* 56, 1247–1293. doi: 10.1007/s10898-012-9951-y

Rodríguez-Fernández, R., Pereira, F. B., Marques, J. M., Martínez-Núñez, E., and Vázquez, S. A. (2017). Gafit: a general-purpose, user-friendly program for fitting potential energy surfaces. *Comput. Phys. Comm.* 217:89. doi: 10.1016/j.cpc.2017.02.008

Rossi, I., and Truhlar, D. G. (1995). Parameterization of NDDO wavefunctions using genetic algorithms. An evolutionary approach to parameterizing potential energy surfaces and direct dynamics calculations for organic reactions. *Chem. Phys. Lett.* 233, 231–236. doi: 10.1016/0009-2614(94)01450-A

Sanz-Sanz, C., Roncero, O., Paniagua, M., and Aguado, A. (2013). Full dimensional potential energy surface for the ground state of $H_4(+)$ system based on triatomic-in-molecules formalism. *J. Chem. Phys.* 139:184302. doi: 10.1063/1.4827640

Shalashilin, D. (2010). Nonadiabatic dynamics with the help of multiconfigurational Ehrenfest method: improved theory and fully quantum 24D simulation of pyrazine. *J. Chem. Phys.* 132:244111. doi: 10.1063/1.3442747

Silva-Junior, M. R., and Thiel, W. (2010). Benchmark of electronically excited states for semiempirical methods: MNDO, AM1, PM3, OM1, OM2, OM3, INDO/S, and INDO/S2. *J. Chem. Theory Comput.* 6:1546. doi: 10.1021/ct100030j

Soto, J., Arenas, J. F., Otero, J. C., and Peláez, D. (2006). Effect of an S_1/S_0 conical intersection on the chemistry of nitramide in its ground state. a comparative CASPT2 study of the nitro-nitrite isomerization reactions in nitramide and nitromethane. *J. Phys. Chem. A* 110:8221. doi: 10.1021/jp0617219

Sparta, M., Hansen, M. B., Matito, E., Toffoli, D., and Christiansen, O. (2010). Using electronic energy derivative information in automated potential energy surface construction for vibrational calculations. *J. Chem. Theory Comput.* 6:3162. doi: 10.1021/ct100229f

Stewart, J. J. (2013). Optimization of parameters for semiempirical methods vi: more modifications to the NDDO approximations and re-optimization of parameters. *J. Mol. Model.* 19:1. doi: 10.1007/s00894-012-1667-x

Stewart, J. J. P. (2016). *Mopac2016, Stewart Computational Chemistry*. Colorado Springs, CO. Available online at: http://OpenMOPAC.net

Thiel, W. (2014). Semiempirical quantum-chemical methods. *WIREs Comput. Mol. Sci.* 4:145. doi: 10.1002/wcms.1161

Toniolo, A., Granucci, G., and Martínez, T. J. (2003). Conical intersections in solution: a QM/MM study using floating occupation semiempirical configuration interaction wave functions. *J. Phys. Chem. A* 107:3822. doi: 10.1021/jp022468p

Troya, D. (2005). Ab initio and direct quasiclassical-trajectory study of the F + CH_4 → HF + CH_3 reaction. *J. Chem. Phys.* 123:214305. doi: 10.1063/1.2126972

Tuckerman, M. E., Marx, D., and Parrinello, M. (2002). The nature and transport mechanism of hydrated hydroxide ions in aqueous solution. *Nature* 417, 925–929. doi: 10.1038/nature00797

Worth, G. A., Beck, M. H., Jäckle, A., and Meyer, H.-D. (2016). *The MCTDH Package, H.-D. Meyer, Version 8.4.12*. Available online at: http://mctdh.uni-hd.de/

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 16 August 2019
doi: 10.3389/fchem.2019.00573

The Structure of Adamantane Clusters: Atomistic vs. Coarse-Grained Predictions From Global Optimization

Javier Hernández-Rojas[1]* *and Florent Calvo*[2]

[1] Departamento de Física e IUdEA, Universidad de La Laguna, San Cristóbal de La Laguna, Spain, [2] Univ. Grenoble Alpes, CNRS, LIPhy, Grenoble, France

Candidate structures for the global minima of adamantane clusters, $(C_{10}H_{16})_N$, are presented. Based on a rigid model for individual molecules with atom-atom pairwise interactions that include Lennard-Jones and Coulomb contributions, low-energy structures were obtained up to $N = 42$ using the basin-hopping method. The results indicate that adamantane clusters initially grow accordingly with an icosahedral packing scheme, followed above $N = 14$ by a structural transition toward face-centered cubic structures. The special stabilities obtained at $N = 13$, 19, and 38 are consistent with these two structural families, and agree with recent mass spectrometry measurements on cationic adamantane clusters. Coarse-graining the intermolecular potential by averaging over all possible orientations only partially confirm the all-atom results, the magic numbers at 13 and 38 being preserved. However, the details near the structural transition are not captured well, because despite their high symmetry the adamantane molecules are still rather anisotropic.

Keywords: global optimization, coarse-grained (CG) model, molecular clusters, potential energy surface (PES), all-atom computer simulations

OPEN ACCESS

Edited by:
Jorge M. C. Marques,
University of Coimbra, Portugal

Reviewed by:
Stefan Goedecker,
University of Basel, Switzerland
Jun Zhang,
Pacific Northwest National Laboratory, United States

****Correspondence:***
Javier Hernández-Rojas
jhrojas@ull.edu.es

Specialty section:
This article was submitted to Theoretical and Computational Chemistry, a section of the journal Frontiers in Chemistry

Received: *29 April 2019*
Accepted: *29 July 2019*
Published: *16 August 2019*

Citation:
Hernández-Rojas J and Calvo F (2019) The Structure of Adamantane Clusters: Atomistic vs. Coarse-Grained Predictions From Global Optimization. Front. Chem. 7:573. doi: 10.3389/fchem.2019.00573

1. INTRODUCTION

Global optimization is an important topic in the physical and chemical sciences, whether we want to refine a force field, predict the native structure of a protein or the crystal structure of some condensed material, or find a practical solution to machine learning problems (Andricioaei and Straub, 1996; Huber and McCammon, 1997; Doye and Wales, 1998; Wales and Hodges, 1998; Wawak et al., 1998; Klepeis and Floudas, 1999; Liwo et al., 1999; Nigra and Kais, 1999; Wales and Scheraga, 1999; Middleton et al., 2001; Hernández-Rojas and Wales, 2003; James et al., 2005; Fadda and Fadda, 2010; Heiles and Johnston, 2013; Wu et al., 2014; Ballard et al., 2016, 2017; Das and Wales, 2016). The case of atomic and molecular clusters is enlightening because such systems exhibit strong finite-size effects, with lowest-energy structures that can depend sensitively and non-monotonically with increasing number of constituents (Stillinger and Weber, 1982, 1984; Tsai and Jordan, 1993b). In particular, efficient global optimization algorithms should be able to explore complex energy landscapes with hierarchical or multifunnel character (Dittes, 1996; Nymeyer et al., 1998; Hamacher and Wenzel, 1999; Wenzel and Hamacher, 1999; Xu and Berne, 1999; Stolovitzky and Berne, 2000; Goedecker, 2004; Cheng et al., 2009; Wang et al., 2010; Oakley et al., 2013).

The difficulties in practically solving the global optimization problem for atomic and molecular systems are at least 2-fold. Firstly, the number of available local minima is thought to increase exponentially with size, making systematic enumeration virtually impossible already above a few tens of particles (Hartke et al., 1998; Wales and Hodges, 1998; Nigra and Kais, 1999; Hodges and Wales, 2000; James et al., 2005; Hernández-Rojas et al., 2006, 2016; Hernández-Rojas and Wales, 2014; Bartolomei et al., 2017). Tsai and Jordan thus evaluated that the 147-atom Lennard-Jones cluster could have more than 10^{60} minima (Tsai and Jordan, 1993a). Secondly, the various structural families generally form different funnels in the landscape separated by high energy barriers, making the sampling problem particularly severe, with conventional simulation methods such as basic molecular dynamics or Monte Carlo, even supplemented with simulated annealing protocols, simply unsuccessful (Wales, 2003).

One additional difficulty arises in molecular systems, even described as rigid bodies, because of the interplay between translational and orientational degrees of freedom. In some cases, the molecules themselves are such that they impose drastic constraints on the collective arrangements that can be adopted by the clusters, starting with the dimer. This occurs, e.g., for planar polycyclic aromatic hydrocarbons (PAHs), which tend to assemble into columnar motifs (Rapacioli et al., 2006; Hernández-Rojas et al., 2016; Bartolomei et al., 2017), or conversely for rodlike molecules, such as CO_2 (Maillet et al., 1998). Even for molecules as relatively simple as water, for which the interactions would seem fairly well-known, water cluster structures are notoriously non-trivial due to the importance and anisotropy of the hydrogen bond (Hartke et al., 1998; Wales and Hodges, 1998; Nigra and Kais, 1999; Hodges and Wales, 2000; James et al., 2005).

In the present work we are interested in clusters of the adamantane molecule ($C_{10}H_{16}$). Adamantane is a small hydrocarbon molecule with pure sp^3 hybridized carbon atoms arranged in a tetrahedral point group, often referred to as a diamondoid. It has a very high thermal stability, and could be found in deep petroleum sources (Dahl et al., 1999, 2010) as well as astrophysical media (Blake et al., 1988; Allamandola et al., 1993; Bauschlicher et al., 2007; Pirali et al., 2008; Steglich et al., 2011). The adamantane molecule is also involved in alkane chemistry (Fokin and Schreiner, 2002), is a versatile building block for larger supramolecular assemblies (Tominaga et al., 2014; Pichierri, 2018) and was found to have some interesting potential in nanomedicine after functionalization (Grillaud et al., 2014; Spilovska et al., 2016; Lee et al., 2018), or even as wheels of nanocars (Chu et al., 2013).

Adamantane clusters were recently synthesized in the cryogenic environment of helium nanodroplets, in which they could be size-selected after ionization by an electron gun (Goulart et al., 2016). In a first approximation, adamantane is roughly spherical and interacts with other molecules via non-covalent forces of the dispersion-repulsion type, with additional Coulomb contributions arising from the partial charges carried by the hydrogen and carbon atoms having different electronegativities. No particular electron delocalization is expected between different molecules, although in the cationic clusters some polarization effects are obviously expected.

So far, the structure of adamantane clusters has not been characterized at the atomistic level of details, but indirect structural information could be drawn from the experimental mass spectra, which show special abundances at the sizes of 13, 19, and 38 molecules. While the former two magic numbers are compatible with icosahedral arrangements, the latter is strongly indicative of a close-packed face-centered cubic structure, suggesting a size-induced structural transition taking place above only a few tens of molecules. Icosahedral-to-cubic transitions are common in atomic and molecular clusters, as they convey the increasing energetic penalty that the highly connected icosahedral structures have to sustain, eventually in favor of less connected but also less strained close-packed structures (Doye et al., 1995; Ikeshoji et al., 2001; Calvo and Carré, 2006). Such a transition has been identified as being strongly influenced by the range of the interparticle potential (Doye et al., 1995; Doye and Wales, 1996). In the present case of adamantane, which has a significant molecular extension while the dispersion interaction is comparatively short-ranged, close packing thus seems natural.

However, the experimental magic numbers do not provide any insight into the orientational ordering within the clusters, and in particular whether the tetrahedral symmetry plays any role on the structures. In order to shed some light onto the relative importance of the translational and orientational degrees of freedom and their interplay, and more generally to confirm whether adamantane clusters do indeed correspond to the speculated structures, we have carried out a systematic global optimization investigation in the size range up to 42 molecules, using the basin-hoping method as our main tool. Two complementary strategies have been employed, namely an all-atom (AA) approach based on a rigid body description, and a highly simplified, coarse-grained (CG) model averaging over all possible orientations.

At the all-atom level, our calculations predict that adamantane clusters are most stable as icosahedra until 14 molecules are reached, and above which the structural arrangement becomes close packed. The special stabilities in the mass spectra are reproduced by the second-energy difference in our all-atom model. At the coarse-grained level, differences appear already above six molecules, although both the icosahedral and cubic motifs at sizes 13 and 38 are correctly reproduced. Comparison between the two models confirms the important role played by the orientational degrees of freedom, despite adamantane being of a rather high symmetry, and shows that the close-packed structures are ideally composed of planes with alternating molecular orientations, a feature that the coarse-grained model is obviously unable to capture.

This paper is organized as follow. We present the potential energy surfaces in section 2 and the methodology employed in the global optimization in section 3. The results are discussed in section 4, and we summarize our conclusions in section 5.

2. POTENTIAL ENERGY SURFACES

Complete global optimization using an explicit description of electronic structure is unfeasible for systems containing hundreds or thousands of atoms, which furthermore can adopt many nearly degenerate local minima. For the present system, and using the model described just below, more than 20 local minima are found just for the adamantane dimer within 2 kJ/mol of the putative global minimum. Moreover, the interactions between neutral adamantane molecules are essentially non-covalent in nature, a notorious issue in quantum chemistry dealing with large molecules. However, the closed-shell electronic structure of the adamantane molecule makes classical force fields particularly attractive for modeling the potential energy surface. A primary assumption usually made at low temperatures relevant for cryogenic environments is to treat the molecules as rigid bodies, with all vibrations frozen. In this work, two models were considered for the interactions between adamantane molecules.

2.1. All-Atom Model

Following the traditional approach of classical force fields, we assume that adamantane molecules interact with each other through a sum of pairwise forces comprising repulsion-dispersion and Coulomb contributions. The interaction V_{ab} between two rigid adamantane molecules a and b is thus expressed by a Lennard-Jones (LJ) part applied between all atoms from a and b, plus electrostatic interactions between partial charges originating from the electronegativity difference between carbon and hydrogen atoms:

$$V_{ab} = \sum_{i,j>i} \left\{ 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j e^2}{4\pi\varepsilon_0 r_{ij}} \right\}, \quad (1)$$

where q_i and q_j are the partial charges on site i of molecule a and site j of molecule b, respectively, r_{ij} is the Cartesian separation between the two sites. In the above expression, all sums were implicitly assumed to be between atoms from different molecules: no intramolecular potential acts for such rigid molecules.

The LJ parameters between sp^3 carbon and hydrogen atoms are taken from the popular OPLS force field (Jorgensen et al., 1996), and read $\varepsilon_{\rm CC} = 0.458$ kJ/mol, $\varepsilon_{\rm HH} = 0.066$ kJ/mol, $\sigma_{\rm CC} = 3.4$ Å, $\sigma_{\rm HH} = 2.649$ Å, Lorentz-Berthelot combination rules providing the complementary values for C-H interactions. The partial charges on individual atoms and the equilibrium geometry of isolated adamantane were obtained from a quantum chemical calculation at the DFT/M06-2X/6-311G(d,p) level of theory. They read $q_{\rm C} = -0.71$ and $q_{\rm H} = +0.14$ for carbon and hydrogen atoms in CH$_2$ groups, and $q_{\rm C} = +0.70$ and $q_{\rm H} = -0.055$ for carbon and hydrogen atoms in CH groups, in units of the electron charge magnitude.

2.2. Comparison With Electronic Structure Calculations

To assess the accuracy and relevance of our simple force field, we have performed dedicated quantum chemical calculations for the adamantane dimer using various levels of theory. Density-functional theory (DFT) is probably the most practical method to deal with such molecules, and here we have chosen the modern functionals PBE0 (Adamo and Barone, 1999), wB97xD (Chai and Head-Gordon, 2008), and M06-2X (Zhao and Truhlar, 2008) as implemented in the Gaussian09 software package (Frisch et al., 2016). While PBE0 does not include explicit dispersion corrections, it performs very well for multipolar descriptions. Both wB97xD and M06-2X are expected to describe non-covalent interactions satisfactorily. Perturbation theory was also employed, using the spin-component-scaled method SCS-MP2 (Grimme, 2003) with basis set superposition errors accounted for using the counterpoise method, as implemented in NWCHEM (Valiev et al., 2010). For these four methods, the two basis sets 6-311G(d,p) and aug-cc-pvDZ were employed independently.

From the resulting geometries, the basic geometric properties of distance r between centers of mass and relative orientations measured by the orientational order parameter κ, as defined below in section 3.2, were evaluated. The interaction energy $E_{\rm int}$ was also determined from the total energies of the optimized monomer and dimer using the standard equation

$$E_{\rm int} = E_{\rm dimer} - 2E_{\rm monomer}, \quad (2)$$

and the resulting values for r, κ and $E_{\rm int}$ are given in **Table 1**.

Unsurprisingly, we find a significant spreading among the DFT results, with a marked dependence of the interaction energy on the functional used, and notably a factor >4 between PBE0 and wB97xD results, SCS-MP2 and M06-2X data lying in-between those extremes. The weaker binding predicted by PBE0 is consistent with this functional not properly accounting for dispersion interactions. Basis set effects further contribute to some variations, although with one magnitude lower. The strong differences between the predictions of PBE0 and wB97xD are comparable to those obtained earlier in other intermolecular interactions problem involving fullerenes and hydrogen (Kaiser et al., 2013; Calvo et al., 2018b).

The force field based on OPLS with multipolar contributions obtained from partial charges derived from DFT performs very satisfactorily against the not-so-extreme quantum chemical predictions from M06-2X and SCS-MP2 both in terms of energy and geometry. The good performance of the force field against the

TABLE 1 | Interaction energy and geometric properties of the adamantane dimer, as predicted by different quantum chemical methods and by the present empirical potential.

Method	$E_{\rm int}$ (kJ/mol)	r (Å)	κ
DFT/PBE0/6-311G(d,p)	−5.19	6.83	−0.31
DFT/PBE0/aug-cc-pvDZ	−7.70	6.59	−0.33
DFT/wB97xD/6-311G(d,p)	−21.33	6.05	−0.25
DFT/wb97xD/aug-cc-pvDZ	−24.72	6.02	−0.26
DFT/M06-2X/6-311G(d,p)	−12.97	6.14	−0.27
DFT/M06-2X/aug-cc-pvDZ	−17.66	6.07	−0.27
SCS-MP2/6-311G(d,p)	−9.97	6.26	−0.31
SCS-MP2/aug-cc-pvDZ	−8.81	6.66	−0.26
Force field	−14.92	6.22	−0.36

The orientational order parameter κ is defined in section 3.2.

Minnesotta functional M06-2X is also consistent with an earlier study on microhydrated RNA precursors (Bacchus-Montabonel and Calvo, 2015) where this quantum chemical method was found to perform better than MP2 against coupled cluster reference data. Together with the difficulty of obtaining more accurate electronic structure properties for the present 52-atom dimer system, these results indicate that our simple model is chemically reliable.

2.3. Coarse-Grained Model

The high symmetry of adamantane encourages us to attempt a simplified description based on a coarse-grained model of the previous all-atom potential. Such an approach has been highly successful in the past for isotropic molecules such as C_{60}, for which simple analytical expressions can be obtained for the integrals (Girifalco, 1991). Here we consider a spherical pointlike version, in which the effective potential is obtained by spherical averaging over the relative orientations of the two molecules, at fixed distance between their centers of mass. Averaging was performed using a random sampling procedure employing 10^6 independent orientational configurations.

In **Figure 1**, the variations of the CG potential are represented against increasing distance, together with the geometry of the equilibrium adamantane dimer obtained at the AA level.

The energy and equilibrium position in the AA model, also highlighted in the figure, show that the CG model underestimates the binding energy by about one order of magnitude, owing to the strong repulsion between peripheral hydrogen atoms, and presents an equilibrium position at a larger distance. The effective potential is very steep, as also expected for an interaction between sizeable molecules. It has thus an effectively short range, which should favor close packing (Doye et al., 1995).

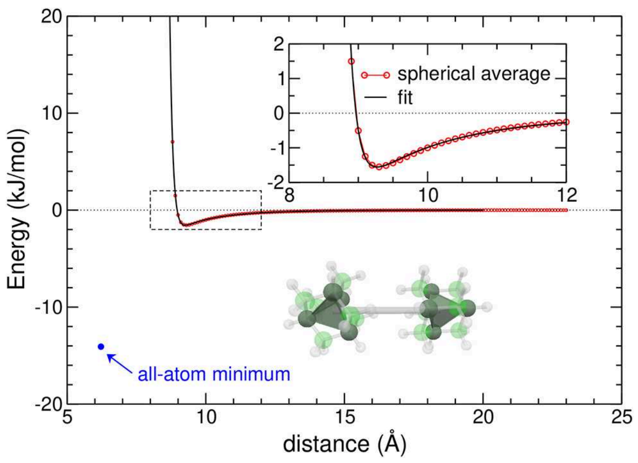


FIGURE 1 | Spherically averaged potential energy curve of the adamantane dimer (red circles) and its best fit giving the coarse-grained potential (black line). The minimum energy in the all-atom model is shown as a blue circle, and the inset highlights the distance range where the potential is minimum. The tetrahedral symmetry of adamantane molecules in the dimer geometry is also shown.

The CG potential can be fitted into a simple expression only dependent on the interparticle distance r as

$$\tilde{V}_{ab} = A\exp[-\alpha(r-r_0)] - f_{\rm cut}\left(\frac{C_6}{r^6}+\frac{C_8}{r^8}\right), \tag{3}$$

with a short-range cut-off function $f_{\rm cut}$ that reads

$$f_{\rm cut} = \begin{cases} \exp\left[-(1-d/r)^2\right] & \text{if } r<d \\ 1 & \text{if } r \geq d \end{cases} \tag{4}$$

The optimal parameters of the CG potential were found to be $A = 0.0468$ kJ/mol, $\alpha = 8.86$ Å^{-1}, $r_0 = 9.405$ Å, $C_6 = 423040.5$ kJ· mol^{-1}·Å^6, $C_8 = 56522581.1$ kJ· mol^{-1}·Å^8, $d = 8$ Å.

3. GLOBAL OPTIMIZATION

The global energy minima were located using the basin-hopping (BH) or Monte Carlo plus minimization method (Li and Scheraga, 1987; Wales and Doye, 1997). The implementation of BH for adamantane clusters differs for the AA and CG potentials due to the presence of orientational degrees of freedom for the former.

3.1. Survey by Basin-Hopping

Basin-hopping is a stochastic algorithm that transforms the PES into a collection of basins of attraction and explore them by random large amplitude, collective moves between minima. This transformed PES preserves all local minima, including the global minima, and the search proceeds by successive applications of the Monte Carlo Metropolis acceptance rule to the locally minimized energies. The BH method has been successfully applied to a plethora of atomic and molecular clusters in the past (Wales et al., 2000; Wales, 2003).

For the CG potential, only translational moves have to be considered, and several series of 10^5 local minimizations were carried out for each cluster size, the fictitious temperature parameter being set such as $k_B T = 0.5$ kJ/mol. No strong influence of this parameter was found here.

For the AA model, the translational and rotational moves can be either managed on a similar footing, or distinguished from one another. In the most general version, a random move thus consists of perturbing all positions of the centers of mass and rotating the molecules, both displacements being performed simultaneously before local minimization is carried out. Here we have chosen to represent the orientational degrees of freedom using angle-axis coordinates $\vec{k} = (n, \ell, m)$, a vector that defines a rotation axis passing through the center of mass and with magnitude of the rotation given by $\theta = \sqrt{n^2 + \ell^2 + m^2}$, relative to a fixed reference frame. This angle-axis representation provides a general framework for rigid body isotropic site-site potentials (Wales, 2005; Chakrabarti and Wales, 2009). The advantage of angle-axis coordinates is that they do not suffer from the so-called gimbal lock problem appearing with Euler angles when rotational axes can become equivalent. Using this framework, the orientational moves consist of perturbing all components of the angle-axis vector producing a new orientation, $\vec{k}' = (n', \ell', m')$, but with the constraint that the new angle $\theta' = \sqrt{n'^2 + \ell'^2 + m'^2}$ remains between 0 and 2π.

Test runs performed for the 12-molecule cluster and employing 5×10^4 BH steps allowed us to evaluate suitable parameters for the basin-hopping optimizations with the AA model, namely $k_{\rm B}T = 1.5$ kJ/mol, giving an acceptance ratio of about 20%. Unfortunately, above size 21 the algorithm was found less efficient, and lower-energy structures could be occasionally found simply by removing molecules from neighboring size clusters and conducting short BH runs. We thus implemented an alternative strategy in which the centers of mass positions were borrowed from the CG minima, purely orientational moves being allowed in the subsequent BH minimization. Here only 10^4 BH collective steps were performed for each cluster size.

The results reported below are thus the results of three combined approaches relying on basin-hopping but altering the entire set of degrees of freedom, only the orientations, or exploring the random removal of one molecule followed by further local search. The orientational minimization was also used to produce all-atom clusters with a specific translational ordering but lying in a different funnel as the global minimum. In practice it allowed us to generate icosahedral and cubic clusters in a broader size range, providing further insight into the related structural transition. In all our BH searches, the geometry was reset to the local minimum before a random perturbation was attempted again.

3.2. Structural Indicators

For the analysis of cluster minima, different order parameters and structural indicators were considered to probe the extent of translational and orientational orderings. The bond-orientational order parameter Q_6 involves the relative positions of the molecular centers of mass and is useful to discriminate icosahedral and cubic packings (Calvo et al., 2018a). It is defined as

$$Q_6 = \left(\frac{4\pi}{13} \sum_{m=-6}^{m=6} \left| \bar{Q}_{6m} \right|^2 \right)^{1/2}, \tag{5}$$

where

$$\bar{Q}_{6m} = \frac{1}{N_b} \sum_{r_{ij} < 7.5\ \text{Å}} Y_{6m}(\theta_{ij}, \phi_{ij}), \tag{6}$$

N_b being the number of bonds defined when the distance between of center of masses of two adamantane molecules is lower than 7.5 Å. $Y_{6m}(\theta_{ij}, \phi_{ij})$ is the spherical harmonic function of degree 6 and order m. The Q_6 parameter can be evaluated for both the AA and CG structures.

An orientational order parameter respecting the tetrahedral symmetry of adamantane was constructed to measure the extent of alignment within the clusters. More precisely, and following Fel (Fel, 1995), for each molecule a we associate four unit vectors $\vec{n}_k^{(a)}$ pointing along the four tetrahedral directions, with Cartesian coordinates $n_{k,\alpha}^{(a)}$ with $\alpha = x, y$, and z. From these coordinates a 3-rank tensor $Q_3^{(a)}$ is constructed as

$$Q_{3,\alpha,\beta,\gamma}^{(a)} = \sum_{k=1}^{4} n_{k\alpha}^{(a)} n_{k\beta}^{(a)} n_{k\gamma}^{(a)}. \tag{7}$$

For a set of molecules, an orientational order parameter κ that is tetrahedrally invariant is defined by considering the pairs of nearest-neighbor molecules as

$$\kappa = \frac{9}{32 n_{nn}} \sum_{a<b, r_{ab}<7.5\ \text{Å}} \text{Tr}\ Q_3^{(a)} Q_3^{(b)}, \tag{8}$$

where n_{nn} is the number of nearest-neighbor molecules. The prefactor ensures that $\kappa = 1$ if all molecules are tetrahedrally aligned.

In addition to purely geometric indicators, energetic parameters were also evaluated to measure the relative stability of the clusters, and quantify the role of orientational strain (*vide infra*).

4. RESULTS

The putative global minima of adamantane clusters were obtained with full atomistic details up to size 42. All structures are available in the **Supplementary Material**. The much less expensive coarse-grained model was able to provide reliable structures in a significantly broader range, although the trends above 42 remain essentially unchanged and will not be discussed specifically.

4.1. Energetic Stability

To estimate the relative stability of different cluster sizes, we evaluated the second-energy derivative of the PES, $\Delta^2 E(N) = E_{N+1} + E_{N-1} - 2E_N$, where E_N is the energy of the global minimum for $(C_{10}H_{16})_N$. Maxima in Δ^2 correspond to clusters with enhanced stability, and are thus closely related to special abundances experimentally measured by mass spectrometry.

The variations of $\Delta^2 E$ with increasing size N are presented in **Figure 2**, as obtained by both the AA and CG models.

From this figure it is clear that the two models do not predict the same special stabilities in the entire size range considered, except at $N = 38$. Prominent peaks in the AA model at $N = 13$, 19, 24, or 29 are not present in the coarse-grained description, and in the range around 30 the differences are rather systematic between the AA and CG models.

The energetic data obtained with atomistic details are essentially consistent with experimental data, indicating that our modeling of adamantane clusters is realistic. The contrasted behaviors between the two models suggest that the mutual orientations of the adamantane molecules play a significant role on the cluster structures.

4.2. Main Structural Motifs

Selected structures obtained with the AA and CG models are presented in **Figure 3**, notably for $N = 13$ and 38 but also

$N = 14$, 15 and $N = 26$, which illustrate the differences between the two descriptions. While the true atomic positions are used for the AA structures, we used fuzzy tetrahedra to represent the adamantane molecules in the CG model.

For $N = 13$, the structure in both models corresponds to an icosahedral packing, however for the CG model the structure does not strictly belong to the I_h point group, the symmetry being lowered due to the important strain within the cluster. In the AA description the molecules manage to adopt appropriate orientations that bring the translational structure closer to the perfect icosahedron.

At size 14 both models predict a qualitatively different structural motif, as a capped icosahedron with all atoms, but showing a decahedral arrangement after coarse-graining. Decahedral motifs are known to occur as an intermediate packing scheme on the way from the highly coordinated, but highly strained icosahedra to the low coordinated and weakly strained close packed structures (Doye et al., 1995). Their presence in the CG model is thus not accidental.

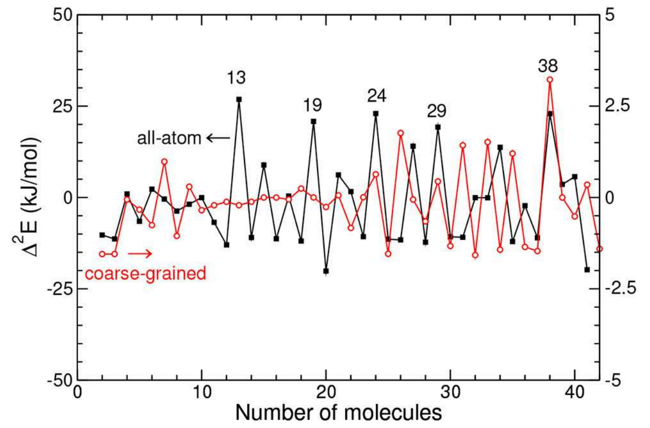


FIGURE 2 | Second-energy derivative vs. cluster size for the all-atom and coarse-grained models.

At size 15 the all-atom model now predicts a cubic motif while the isotropic potential still yields a (doubly capped) decahedron. The cubic translational arrangement is preserved at sizes 16 and beyond, while the coarse-grained model further experiences some structural changes. At sizes 26 and above, both models favor close-packed cubic structures, leading to the perfect truncated octahedron at $N = 38$ as a strong magic number. These results thus support the interpretation of experimental mass spectra from the Scheier group (Goulart et al., 2016), namely that adamantane clusters exhibit icosahedral and cubic packing as their main structural motifs, at low and large sizes, respectively. Our results indicate that icosahedral packing is the dominant motif only up to $N = 14$, and that orientational effects are already non-negligible at this size.

4.3. Structural Analysis

To shed more light onto the respective roles of translational and orientational orderings on the stable structures of adamantane clusters, and to clarify the effects of coarse-graining, we now consider the structural order parameters introduced in section 3.2 in comparison between the two models. Near size 14 where the icosahedral-cubic transition takes place, additional but metastable structures were generated as belonging to the icosahedral and cubic families, by performing basin-hopping global optimization with orientational moves only.

The bond-orientational order parameter Q_6 is shown against increasing cluster size for both models in **Figure 4**.

Within the all-atom description, Q_6 exhibits irregular, essentially decreasing variations during the completion of

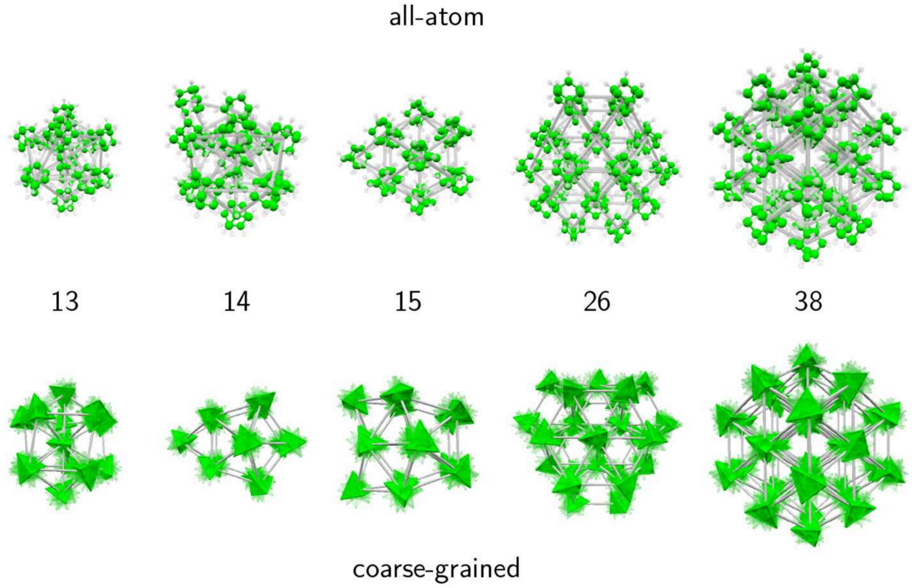


FIGURE 3 | Remarkable structures obtained for selected adamantane clusters with 13–15, 26, and 38 molecules, in the all-atom and coarse-grained models.

icosahedral packing at $N = 13$. Above this size, Q_6 reaches about 0.58 and stays constant at this value, indicating that the face-centered cubic structure is robust and regular with no point defect or stacking fault.

In the CG model, Q_6 displays the same value as in the AA description up to size 7, indicating that translational structures are identical. Differences above the critical size of $N = 14$ show that the cubic packing is less ideal for the CG model, except near size 40 where Q_6 reaches the same value as in the AA model. As confirmed by visual inspection along the lines of **Figure 3**, decahedral packings are often found, with a signature on Q_6 being lower than ∼0.4, except for $N = 24$–28, $N = 34$, and $N > 36$ for which the cubic motif is lower in energy, Q_6 being also higher.

The differences between the AA and CG models further support that orientational ordering plays a role in establishing the close-packed translational ordering itself. To further explore this aspect, the order parameter κ was evaluated for atomistic global minima, the results of which are depicted in **Figure 5** against cluster size.

Similar to Q_6, the orientational order parameter displays irregular variations during the completion of the icosahedron at $N = 13$, with positive and negative values alike. The tetrahedra in this size range thus do not possess any robust and specific orientational preference. Once the cubic packing is set at $N \geq 15$, and as was the case for the translational order parameter, κ reaches an essentially constant value close to -0.25, with fluctuations of magnitude no greater than 0.05.

This stability further indicates that the clusters adopt a constant growth scheme. However, the negative value of κ also shows that the tetrahedral molecules do not display a single, common orientation within the cluster, as otherwise κ would be closer to unity. To illustrate the specific orientational ordering, we have represented in **Figure 6** another set of clusters obtained for both the AA and CG models, and chosen at sizes for which the AA description predicts special stabilities, namely 19, 24, and 29. In the AA model, the atomic details were replaced by equivalent tetrahedra, contrasting with the fuzzy tetrahedra in the CG model.

In this size range, the coarse-grained potential predicts both decahedral (19 and 29) and cubic (24) motifs. The structures obtained with the AA model show the same close-packed motif, with clusters of a given size that are subparts of larger global minima. More interestingly, and as suggested by the indicators previously discussed, the molecules show two different possible orientations that alternate between planes in the largest

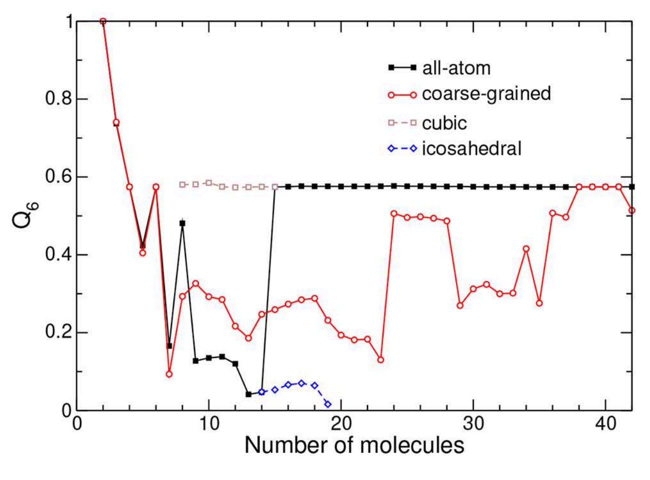


FIGURE 4 | Bond-orientational order parameter Q_6 obtained from the relative positions of the centers of mass of adamantane clusters in the all-atom and coarse-grained models. The values obtained for metastable icosahedral and cubic conformations near the corresponding transition are also shown.

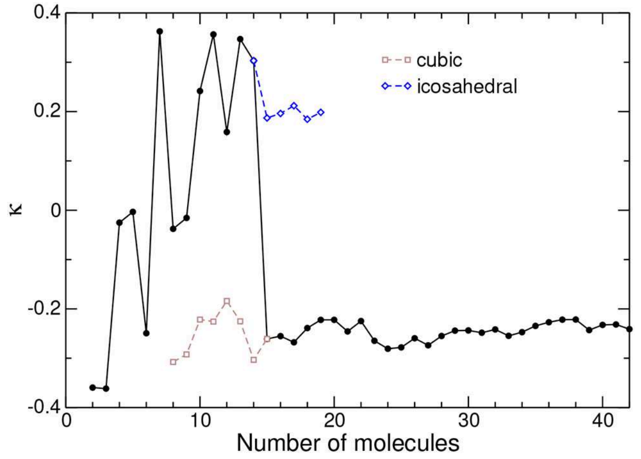


FIGURE 5 | Tetrahedral order parameter κ between nearest-neighbor molecules in adamantane clusters, as obtained from the all-atom model. The values obtained for metastable icosahedral and cubic conformations near the corresponding transition are also shown.

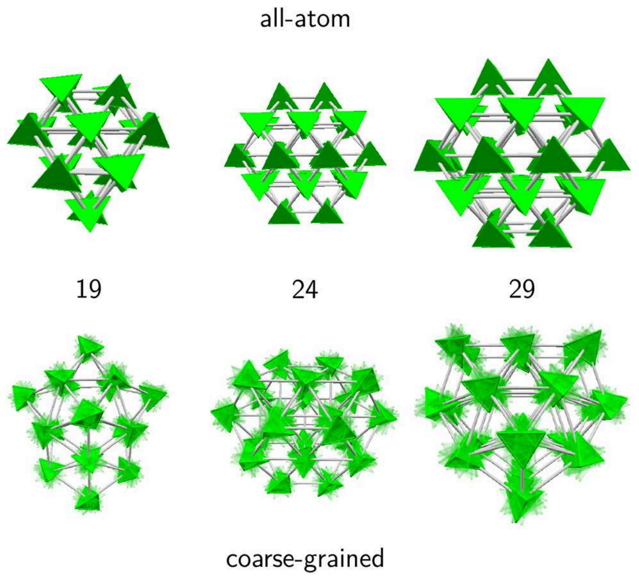


FIGURE 6 | Selected adamantane clusters for which the second energy derivative shows peaks in the 19–29 size range. In the all-atom model, molecules were replaced by their equivalent tetrahedra to emphasize orientational ordering.

clusters. While molecules with the same orientation within a same plane are next nearest neighbors, nearest-neighbor molecules precisely belong to different planes and present parallel contact faces.

However, in the dimer at equilibrium (see **Figure 1**), the two tetrahedra do not display such a relative orientation, and instead rotate in order to maximize dispersive attractions while minimizing Coulomb repulsion between the (positively charged) peripheral hydrogens. In clusters, this difference in relative orientations gives rise to orientational strain (Calvo et al., 1999), which the system exploits to minimize the overall energy while deviating from the ideal orientations that would be adopted in absence of environment.

We have quantified the importance of strain in adamantane clusters by removing from their total potential energy the contribution between nearest-neighbor pairs, as if these pairs were at equilibrium (including their orientational degrees of freedom) (Doye et al., 1995). Omitting the contribution of non-nearest neighbors, the strain energy V_{strain} reads

$$V_{\mathrm{strain}} = \sum_{a<b}^{nn \text{ only}} V_{ab} - n_{nn} V_{\min}, \qquad (9)$$

where $V_{\min}$ denotes the minimum energy in the dimer at equilibrium.

In order to compare the two models, we have deemed more suitable to further normalize the strain energy by the magnitude of the dimer binding energy, considering thus a strain factor $V_{\mathrm{strain}}/|V_{\min}|$ instead of the absolute strain energy. The variations of the strain factor with increasing size are represented in **Figure 7** for both models.

With the coarse-grained description, which ignores orientational degrees of freedom, most structures are either icosahedral or decahedral and thus exhibit moderate strain (Doye et al., 1995), cubic packings being characterized with a very low strain factor. In this respect, the strain factor is an even more direct probe of close packed structures than Q_6 previously considered.

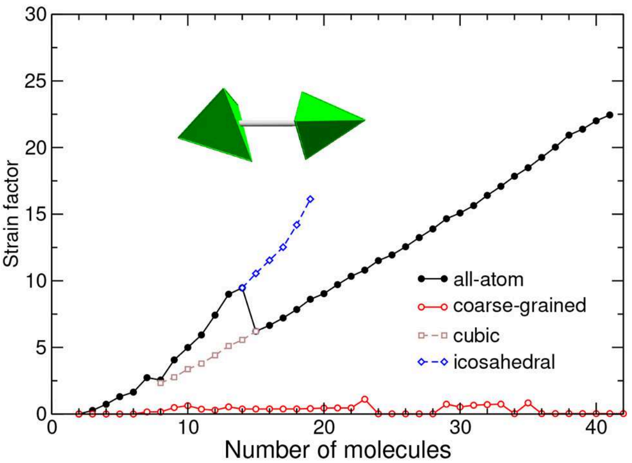


FIGURE 7 | Strain energy normalized by dimer energy, as a function of cluster size and for the all-atom and coarse-grained models. The equivalent tetrahedra in the equilibrium dimer geometry are depicted. The values obtained for metastable icosahedral and cubic conformations near the corresponding transition are also shown.

In contrast, the all-atom model shows strongly increasing strain as the cluster size increases, with a peak at $N = 14$ and a change in slope above this size. The growing strain conveys the inability of the adamantane molecules to respect their ideal mutual orientation in the equilibrium dimer. However, in fairness it should be recognized that this orientation is not so meaningful as soon as the cubic motif is established. If instead of the equilibrium dimer we had artificially chosen the orientations between nearest neighbor in close-packed clusters to define the strain energy, the strain factor would be much reduced and similar to the value in the CG model, but the values in icosahedral structures would become negative and less physical.

5. CONCLUDING REMARKS

The remarkable thermodynamical and chemical stability of adamantane makes it a valuable building block of supramolecular materials, including non-covalent molecular clusters. Recent mass spectrometry measurements under the cryogenic conditions of helium droplets have found magic numbers for cationic adamantane clusters at the sizes of 13, 19, and 38 molecules, as well as others higher suggesting close packed geometries (Goulart et al., 2016). In the present work, we have modeled (neutral) adamantane clusters using a rigid body description and a site-site pairwise force field comprising the traditional Lennard-Jones potential for repulsion-dispersion forces with Coulomb interactions acting between partial charges. A spherically averaged coarse-grained model was also developed, producing an effective pair potential that allows an efficient exploration of the translational structure of adamantane clusters. The all-atom force field was successfully validated against quantum chemistry calculations, which incidentally highlighted the difficulty of producing accurate and reliable interaction energies and geometries for such rather large non-covalent edifices.

Using the basin-hopping algorithm, the putative global minima of $(C_{10}H_{16})_N$ clusters were found to follow icosahedral packing up to $N = 14$ and sharply change into close-packed cubic structure above this size. Translational and orientational order parameters indicate that cubic structures are stabilized by having molecules with two possible orientations in alternating planes. This feature is obviously absent with the CG model, which predicts numerous decahedral structures in the intermediate range 14–35, before the structure eventually also adopts the close-packed cubic motif; this intermediate decahedral phase is absent from the all-atom structures.

Comparison between the all-atom and coarse-grained models highlights and explains the importance of orientational strain in the structure of adamantane clusters, in particular the sharp transition toward the cubic motif which arises due to a combination between the short range of the potential and the optimal orientations presented by the nearest-neighbor molecules with tetrahedral facets parallel to one another.

Here we have neglected the cationic nature of the adamantane clusters in the mass spectrometry experiments, but in a first approximation it could be accounted for by adding a polarization contribution and assuming the N-molecule cationic cluster to be made of a single cationic molecule surrounded by $N-1$ neutral ones. Such an additional contribution would bind the first solvation shell more strongly, possibly leading to some structural distortion, and could even modify the details of the icosahedral-to-cubic transition, but would probably not change the qualitative picture or the special stabilities found at 13 or 38. Further efforts should also be devoted to making the basin-hopping optimization method even more efficient for the present clusters. Although we have focused on the chemical physics rather than the algorithmic efficiency, it was clear that basin-hopping in its conventional approach was struggling to locate the correct molecular orientations in medium- to large-size clusters. Having analyzed the structures, such a deceiving efficiency appears more clearly and is most likely due to the collective nature of the orientational ordering in the clusters, where the orientation is constant within a plane but alternates between planes. Tailored moves that incorporate such a specificity should enable much larger clusters to be addressed in the future.

DATA AVAILABILITY

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

JH-R and FC conceived the project, performed the electronic structure calculations and prepared, wrote, and discussed the manuscript. FC built and conducted the coarse-grained calculations. JH-R conducted the all-atom basin-hopping calculations.

FUNDING

JH-R acknowledges the financial support from Ministerio de Economía y Competitividad (Spain) under Grant No. FIS2016-79596-P.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2019.00573/full#supplementary-material

REFERENCES

Adamo, C., and Barone, V. (1999). Toward reliable density functional methods without adjustable parameters: the PBE0 model. *J. Chem. Phys.* 110:6158. doi: 10.1063/1.478522

Allamandola, L., Sandford, S., Tielens, A., and Herbst, T. (1993). Diamonds in dense molecular clouds: a challenge to the standard interstellar medium paradigm. *Science* 260, 64–66. doi: 10.1126/science.11538059

Andricioaei, I., and Straub, J. E. (1996). Generalized simulated annealing algorithms using Tsallis statistics: application to conformational optimization of a tetrapeptide. *Phys. Rev. E* 53:R3055. doi: 10.1103/PhysRevE.53.R3055

Bacchus-Montabonel, M. -C., and Calvo, F. (2015). Nanohydration of uracil: emergence of three-dimensional structures and proton-induced charge transfer. *Phys. Chem. Chem. Phys.* 17, 9629–9633. doi: 10.1039/C5CP00611B

Ballard, A. J., Das, R., Martiniani, S., Mehta, D., Sagun, L., Stevenson, J. D., et al. (2017). Energy landscapes for machine learning. *Phys. Chem. Chem. Phys.* 19, 12585–12603. doi: 10.1039/C7CP01108C

Ballard, A. J., Stevenson, J. D., Das, R., and Wales, D. J. (2016). Energy landscapes for a machine learning application to series data. *J. Chem. Phys.* 144:124119. doi: 10.1063/1.4944672

Bartolomei, M., Pirani, F., and Marques, J. M. C. (2017). Modeling coronene nanostructures: analytical potential, stable configurations and *ab initio* energies. *J. Phys. Chem. C* 121, 14330–14338. doi: 10.1021/acs.jpcc.7b03691

Bauschlicher, C., Liu, Y., Ricca, A., Mattioda, A., and Allamandola, L. (2007). Electronic and vibrational spectroscopy of diamondoids and the interstellar infrared bands between 3.35 and 3.55 μm. *Astrophys. J.* 671, 458–469. doi: 10.1086/522683

Blake, D., Freund, F., Krishnan, K., Echer, C. J., Shipp, R., Bunch, T. E., et al. (1988). The nature and origin of interstellar diamond. *Nature* 332, 611–613. doi: 10.1038/332611a0

Calvo, F., Boutin, A., and Labastie, P. (1999). Structure of nitrogen molecular clusters $(N_2)_n$ with $13 \leq n \leq 55$. *Eur. Phys. J. D* 9, 189–193. doi: 10.1007/978-3-642-88188-6-37

Calvo, F., and Carré, A. (2006). Structural transitions and stabilization of palladium nanoparticles upon hydrogenation. *Nanotechnology* 17, 1292–1299. doi: 10.1088/0957-4484/17/5/022

Calvo, F., Hamdi, R., Mejrissi, L., and Oujia, B. (2018a). Questioning the structure of $Sr^+(Ar)_n$ clusters. *Eur. Phys. J. D* 72:133. doi: 10.1140/epjd/e2018-90160-5

Calvo, F., Yurtsever, E., and Tekin, A. (2018b). Physisorption of H_2 on fullerenes and the solvation of C_{60} by hydrogen clusters at finite temperature: a theoretical assessment. *J. Phys. Chem. A* 122, 2792–2800. doi: 10.1021/acs.jpca.8b00163

Chai, J. -D., and Head-Gordon, M. (2008). Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.* 10:6615. doi: 10.1039/b810189b

Chakrabarti, D., and Wales, D. J. (2009). Simulations of rigid bodies in an angle-axis framework. *Phys. Chem. Chem. Phys.* 11, 1970–1976. doi: 10.1039/b818054g

Cheng, L., Feng, Y., Yang, J., and Yang, J. (2009). Funnel hopping: searching the cluster potential energy surface over the funnels. *J. Chem. Phys.* 130:214112. doi: 10.1063/1.3152121

Chu, P. L. E., Wang, L. Y., Khatua, S., Kolomeisky, A. B., Link, S., and Tour, J. M. (2013). Synthesis and single-molecule imaging of highly mobile adamantane-wheeled nanocars. *ACS Nano* 7, 35–41. doi: 10.1021/nn304584a

Dahl, J. E., Moldowan, J. M., Peters, K. E., Claypool, G. E., Rooney, M. A., Michael, G. E., et al. (1999). Diamondoid hydrocarbons as indicators of natural oil cracking. *Nature* 399, 54–57. doi: 10.1038/19953

Dahl, J. E., Moldowan, J. M., Wei, Z., Lipton, P. A., Denisevich, P., Gat, R., et al. (2010). Synthesis of higher diamondoids and implications for their formation in petroleum. *Angew. Chem. Int. Ed.* 49, 9881–9885. doi: 10.1002/anie.201004276

Das, R., and Wales, D. J. (2016). Energy landscapes for a machine-learning prediction of patient discharge. *Phys. Rev. E* 93:063310. doi: 10.1103/PhysRevE.93.063310

Dittes, F.-M. (1996). Optimization on rugged landscapes: a new general purpose Monte Carlo approach. *Phys. Rev. Lett.* 76:4651. doi: 10.1103/PhysRevLett.76.4651

Doye, J. P. K., and Wales, D. J. (1996). The effect of the range of the potential on the structure and stability of simple liquids: from clusters to bulk, from sodium to. *J. Phys. B* 29, 4859–4894. doi: 10.1088/0953-4075/29/21/002

Doye, J. P. K., and Wales, D. J. (1998). Global minima for transition metal clusters described by Sutton-Chen potentials. *New J. Chem.* 22, 733–744. doi: 10.1039/a709249k

Doye, J. P. K., Wales, D. J., and Berry, R. S. (1995). The effect of the range of the potential on the structures of clusters. *J. Chem. Phys.* 103, 4234–4249. doi: 10.1063/1.470729

Fadda, A., and Fadda, G. (2010). An evolutionary algorithm for the prediction of crystal structures. *Phys. Rev. B* 82:104105. doi: 10.1103/PhysRevB.82.104105

Fel, L. G. (1995). Tetrahedral symmetry in nematic liquid crystals. *Phys. Rev. E* 52, 702–717. doi: 10.1103/PhysRevE.52.702

Fokin, A. A., and Schreiner, P. R. (2002). Selective alkane transformations via radicals and radical cations: insights into the activation step from experiment and theory. *Chem. Rev.* 102, 1551–1593. doi: 10.1021/cr000453m

Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., et al. (2016). *Gaussian 09 Revision D.01.* Gaussian Inc., Wallingford, CT 2009.

Girifalco, L. A. (1991). Interaction potential for carbon (C60) molecules. *J. Phys. Chem*. 95, 5370–5371. doi: 10.1021/j100167a002

Goedecker, S. (2004). Minima hopping: an efficient search method for the global minimum of the potential energy surface of complex molecular systems. *J. Chem. Phys.* 120, 9911–9917. doi: 10.1063/1.1724816

Goulart, M., Kuhn, M., Kranabetter, L., Kaiser, A., Postler, J., Rastogi, M., et al. (2016). Magic numbers for packing adamantane in helium droplets: cluster cations, dications and trications. *J. Phys. Chem. C* 121, 10767–10772. doi: 10.1021/acs.jpcc.6b11330

Grillaud, M., Russier, J., and Bianco, A. (2014). Polycationic adamantane-based dendrons of different generations display high cellular uptake without triggering cytotoxicity. *J. Am. Chem. Soc.* 136, 810–819. doi: 10.1021/ja411987g

Grimme, S. (2003). Improved second-order Møller-Plesset perturbation theory by separate scaling of parallel- and antiparallel-spin pair correlation energies. *J. Chem. Phys.* 118, 9095–9102. doi: 10.1063/1.1569242

Hamacher, K., and Wenzel, W. (1999). Scaling behavior of stochastic minimization algorithms in a perfect funnel landscape. *Phys. Rev. E* 59, 938–941. doi: 10.1103/PhysRevE.59.938

Hartke, B., Schutz, M., and Werner, H. J. (1998). Improved intermolecular water potential from global geometry optimization of small water clusters using local MP2. *Chem. Phys.* 239, 561–572. doi: 10.1016/S0301-0104(98)00322-X

Heiles, S., and Johnston, R. L. (2013). Global optimization of clusters using electronic structure methods. *Int. J. Quantum Chem.* 113, 2091–2109. doi: 10.1002/qua.24462

Hernández-Rojas, J., Breton, J., Llorente, J. M. G., and Wales, D. J. (2006). Global potential energy minima of $C_{60}(H_2O)_n$ clusters. *J. Phys. Chem. B* 110, 13357–13362. doi: 10.1021/jp0572582

Hernández-Rojas, J., Calvo, F., and Wales, D. J. (2016). Coarse-graining the structure of polycyclic aromatic hydrocarbons clusters. *Phys. Chem. Chem. Phys.* 18, 13736–13740. doi: 10.1039/C6CP00592F

Hernández-Rojas, J., and Wales, D. J. (2003). Global minima for rare gas clusters containing one alkali metal ion. *J. Chem. Phys.* 119, 7800–7804. doi: 10.1063/1.1608852

Hernández-Rojas, J., and Wales, D. J. (2014). The effect of dispersion damping functions on the structure of water clusters. *Chem. Phys.* 444, 23–29. doi: 10.1016/j.chemphys.2014.09.013

Hodges, M. P., and Wales, D. J. (2000). Global minima of protonated water clusters. *Chem. Phys. Lett.* 324, 279–288. doi: 10.1016/S0009-2614(00)00584-4

Huber, G. A., and McCammon, J. A. (1997). Weighted-ensemble simulated annealing: faster optimization on hierarchical energy surfaces. *Phys. Rev. E* 55:4822. doi: 10.1103/PhysRevE.55.4822

Ikeshoji, T., Torchet, G., de Feraudy, M. -F., and Koga, K. (2001). Icosahedron-fcc transition size by molecular dynamics simulation of Lennard-Jones clusters at a finite temperature. *Phys. Rev. E* 63:031101. doi: 10.1103/PhysRevE.63.031101

James, T., Wales, D. J., and Hernández-Rojas, J. (2005). Global minima for water clusters $(H_2O)_n$, $n < 21$, described by a five-site empirical potential. *Chem. Phys. Lett.* 415, 302–307. doi: 10.1016/j.cplett.2005.09.019

Jorgensen, W. L., Maxwell, D. S., and Tirado-Rives, J. (1996). Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* 118, 11225–11236. doi: 10.1021/ja9621760

Kaiser, A., Leidlmair, C., Bartl, P., Zöttl, S., Denifl, S., Mauracher, A., et al. (2013). Adsorption of hydrogen on neutral and charged fullerene: experiment and theory. *J. Chem. Phys.* 138:074311. doi: 10.1063/1.4790403

Klepeis, J. L., and Floudas, C. A. (1999). Free energy calculations for peptides via deterministic global optimization. *J. Chem. Phys.* 110, 7491–7512. doi: 10.1063/1.478652

Lee, D. W., Jo, J., Jo, D., Kim, J., Min, J. J., Yang, D. H., et al. (2018). Supramolecular assembly based on host-guest interaction between beta-cyclodextrin and adamantane for specifically targeted cancer imaging. *J. Indust. Engin. Chem.* 57, 37–44. doi: 10.1016/j.jiec.2017.08.005

Li, Z., and Scheraga, H. A. (1987). Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc. Natl. Acad. Sci. U.S.A.* 84, 6611–6615. doi: 10.1073/pnas.84.19.6611

Liwo, A., Lee, J., Ripoll, D. R., Pillardy, J., and Scheraga, H. A. (1999). Protein structure prediction by global optimization of a potential energy function. *Proc. Natl. Acad. Sci. U.S.A.* 96, 5482–5485. doi: 10.1073/pnas.96.10.5482

Maillet, J.-B., Boutin, A., Buttefey, S., Calvo, F., and Fuchs, A. H. (1998). From molecular clusters to bulk matter. I. Structure and thermodynamics of small CO_2, N_2, and SF_6 clusters. *J. Chem. Phys.* 109, 329–338. doi: 10.1063/1.476509

Middleton, T. F., Hernández-Rojas, J., Mortenson, P. N., and Wales, D. J. (2001). Crystals of binary Lennard-Jones solids. *Phys. Rev. B* 64:184201. doi: 10.1103/PhysRevB.64.184201

Nigra, P., and Kais, S. (1999). Pivot method for global optimization: a study of water clusters $(H_2O)_n$ with $2 \leq n \leq 33$. *Chem. Phys. Lett.* 305, 433–438. doi: 10.1016/S0009-2614(99)00423-6

Nymeyer, H., García, A. E., and Onuchic, J. N. (1998). Folding funnels and frustration in off-lattice minimalist protein landscapes. *Proc. Natl. Acad. Sci. U.S.A.* 95:5921. doi: 10.1073/pnas.95.11.5921

Oakley, M. T., Johnston, R. L., and Wales, D. J. (2013). Symmetrisation schemes for global optimisation of atomic clusters. *Phys. Chem. Chem. Phys.* 15, 3965–3976. doi: 10.1039/c3cp44332a

Pichierri, F. (2018). Adamantane template effect on the self-assembly of a molecular tetrahedron: a theoretical analysis. *Chem. Phys. Lett.* 713, 149–152. doi: 10.1016/j.cplett.2018.10.032

Pirali, O., Vervloet, M., Dahl, J. E., Carlson, R., Tielens, A. G. G. M., and Oomens, O. (2008). Infrared spectroscopy of diamondoid molecules: new insights into the presence of nanodiamonds in the interstellar medium. *Astrophys. J.* 661:919. doi: 10.1086/516731

Rapacioli, M., Calvo, F., Joblin, C., Parneix, P., Toublanc, D., and Spiegelman, F. (2006). Formation and destruction of polycyclic aromatic hydrocarbon clusters in the interstellar medium. *Astron. Astrophys.* 460, 519–531. doi: 10.1051/0004-6361:20065412

Spilovska, K., Zemek, F., Korabecny, J., Nepovimova, E., Soukup, O., Windisch, M., et al. (2016). Adamantane - a lead structure for drugs in clinical practice. *Curr. Med. Chem.* 23, 3245–3266. doi: 10.2174/0929867323666160525114026

Steglich, M., Huisken, F., Dahl, J. E., Carlson, R., and Henning, T. (2011). Electronic spectroscopy of FUV-irradiated diamondoids: a combined experimental and theoretical study. *Astrophys. J.* 729:91. doi: 10.1088/0004-637X/729/2/91

Stillinger, F. H., and Weber, T. A. (1982). Hidden structure in liquids. *Phys. Rev. A* 25:978. doi: 10.1103/PhysRevA.25.978

Stillinger, F. H., and Weber, T. A. (1984). Packing structures and transitions in liquids and solids. *Science* 225:983. doi: 10.1126/science.225.4666.983

Stolovitzky, G., and Berne, B. J. (2000). Catalytic tempering: a method for sampling rough energy landscapes by Monte Carlo. *Proc. Nat. Acad. Sci. U.S.A.* 97, 11164–11169. doi: 10.1073/pnas.97.21.11164

Tominaga, M., Ukai, H., Katagiri, K., Ohara, K., Yamaguchi, K., and Azumaya, I. (2014). Tubular structures bearing channels in organic crystals composed of adamantane-based macrocycles. *Tetrahedron* 70, 2576–2581. doi: 10.1016/j.tet.2014.02.006

Tsai, C. J., and Jordan, K. D. (1993a). Use of an eigenmode method to locate the stationary points on the potential energy surfaces of selected argon and water clusters. *J. Phys. Chem.* 97, 11227–11237. doi: 10.1021/j100145a019

Tsai, C. J., and Jordan, K. D. (1993b). Use of the histogram and jump-walking methods for overcoming slow barrier crossing behaviour in Monte Carlo simulations: applications to the phase transitions in the Ar_{13} and $(H_2O)_8$ clusters. *J. Chem. Phys.* 99:6957. doi: 10.1063/1.465442

Valiev, M., Bylaska, E. J., Govind, N., Kowalski, K., Straatsma, T. P., van Dam, H. J. J., et al. (2010). NWChem: a comprehensive and scalable open-source solution for large scale molecular simulations. *Comput. Phys. Commun.* 181:1477. doi: 10.1016/j.cpc.2010.04.018

Wales, D. J. (2003). *Energy Landscapes*. Cambridge: Cambridge University Press.

Wales, D. J. (2005). The energy landscape as a unifying theme in molecular science. *Phil. Trans. Roy. Soc. A* 363, 357–377. doi: 10.1098/rsta.2004.1497

Wales, D. J., and Doye, J. P. K. (1997). Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *J. Phys. Chem. A* 111, 5111–5116. doi: 10.1021/jp970984n

Wales, D. J., Doye, J. P. K., Miller, M. A., Mortenson, P. A., and Walsh, T. R. (2000). Energy landscapes: from clusters to biomolecules. *Adv. Chem. Phys.* 115, 1–111. doi: 10.1002/9780470141748.ch1

Wales, D. J., and Hodges, M. P. (1998). Global minima of water clusters $(H_2O)_n$, $n \leq 21$, described by an empirical potential. *Chem. Phys. Lett.* 286, 65–72. doi: 10.1016/S0009-2614(98)00065-7

Wales, D. J., and Scheraga, H. A. (1999). Global optimization of clusters, crystals and biomolecules. *Science* 285, 1368–1372. doi: 10.1126/science.285.5432.1368

Wang, Y., Lv, J., Zhu, L., and Ma, Y. (2010). Crystal structure prediction via particle-swarm optimization. *Phys. Rev. B* 82:094116. doi: 10.1103/PhysRevB.82.094116

Wawak, R. J., Pillardy, J., Liwo, A., Gibson, K. D., and Scheraga, H. A. (1998). Diffusion equation and distance scaling methods of global optimization: applications to crystal structure prediction. *J. Phys. Chem. A* 102, 2904–2918. doi: 10.1021/jp972424u

Wenzel, W., and Hamacher, K. (1999). Stochastic tunneling approach for global minimization of complex potential energy landscapes. *Phys. Rev. Lett.* 82, 3003–3007. doi: 10.1103/PhysRevLett.82.3003

Wu, S. Q., Ji, M., Wang, C. Z., Nguyen, M. C., Zhao, X., Umemoto, K., et al. (2014). An adaptive genetic algorithm for crystal structure prediction. *J. Phys.* 26:035402. doi: 10.1088/0953-8984/26/3/035402

Xu, H., and Berne, B. J. (1999). Multicanonical jump walking: a method for efficiently sampling rough energy landscapes. *J. Chem. Phys.* 110, 10299–10306. doi: 10.1063/1.478963

Zhao, Y., and Truhlar, D. G. (2008). The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* 120, 215–241. doi: 10.1007/s00214-007-0310-x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Chemistry

ORIGINAL RESEARCH
published: 12 September 2019
doi: 10.3389/fchem.2019.00624

Hydrated Sodium Ion Clusters $[Na^+(H_2O)_n$ (n = 1–6)]: An *ab initio* Study on Structures and Non-covalent Interaction

Pengju Wang [1], ***Ruili Shi*** [1,2]*, ***Yan Su*** [1]*, ***Lingli Tang*** [3], ***Xiaoming Huang*** [4] ***and Jijun Zhao*** [1]

[1] Key Laboratory of Materials Modification by Laser, Ion and Electron Beams (Dalian University of Technology), Ministry of Education, Dalian, China, [2] School of Mathematics and Physics, Hebei University of Engineering, Handan, China, [3] College of Science, Dalian Nationalities University, Dalian, China, [4] School of Ocean Science and Technology, Dalian University of Technology, Panjin, China

Structural, thermodynamic, and vibrational characteristics of water clusters up to six water molecules incorporating a single sodium ion $[Na^+(H_2O)_n$ (n = 1–6)] are calculated using a comprehensive genetic algorithm combined with density functional theory on global search, followed by high-level *ab initio* calculation. For $n \geq 4$, the coordinated water molecules number for the global minimum of clusters is 4 and the outer water molecules connecting with coordinated water molecules by hydrogen bonds. The charge analysis reveals the electron transfer between sodium ions and water molecules, providing an insight into the variations of properties of O–H bonds in clusters. Moreover, the simulated infrared (IR) spectra with anharmonic correction are in good agreement with the experimental results. The O–H stretching vibration frequencies show redshifts comparing with a free water molecule, which is attributed to the non-covalent interactions, including the ion–water interaction, and hydrogen bonds. Our results exhibit the comprehensive geometries, energies, charge, and anharmonic vibrational properties of $Na^+(H_2O)_n$ (n = 1–6), and reveal a deeper insight of non-covalent interactions.

Keywords: hydrated sodium cluster, stabilization energy, anharmonic effect, IR spectra, natural bond orbital

OPEN ACCESS

Edited by:
Jorge M. C. Marques,
University of Coimbra, Portugal

Reviewed by:
Bernd Hartke,
University of Kiel, Germany
James M. Lisy,
University of Illinois at
Urbana-Champaign, United States

****Correspondence:***
Ruili Shi
ruili_shi@mail.dlut.edu.cn
Yan Su
su.yan@dlut.edu.cn

Specialty section:
This article was submitted to
Theoretical and Computational
Chemistry,
a section of the journal
Frontiers in Chemistry

Received: *30 April 2019*
Accepted: *29 August 2019*
Published: *12 September 2019*

Citation:
Wang P, Shi R, Su Y, Tang L, Huang X and Zhao J (2019) Hydrated Sodium Ion Clusters $[Na^+(H_2O)_n$ (n = 1–6)]: An ab initio Study on Structures and Non-covalent Interaction. Front. Chem. 7:624. doi: 10.3389/fchem.2019.00624

INTRODUCTION

Hydrated ion clusters widely exist in oceans and living organisms, especially hydrated sodium ion clusters, which are important in the control of blood pressure, cell permeability, neuronal activity, and other somatic functions (Jensen, 1992; Feller et al., 1994; Pohl et al., 2013). Understanding the behavior of hydrated sodium ion clusters is helpful to uncover the mechanism of some key biochemical reactions (Mano and Driscoll, 1999; Snyder, 2002; Dudev and Lim, 2010; Payandeh et al., 2011). A number of experimental (Dzidic and Kebarle, 1970; Tang and Castleman, 1972; Schulz et al., 1986, 1988; Blades et al., 1990; Hertel et al., 1991; Patwari and Lisy, 2003; Vaden et al., 2004; Mancinelli et al., 2007) and theoretical (Perez et al., 1983; Arbman et al., 1985; Lybrand and Kollman, 1985; Cieplak et al., 1987; Probst, 1987; Bauschlicher et al., 1991; Dang et al., 1991; Perera and Berkowitz, 1991; Hashimoto and Morokuma, 1994; Glendening and Feller, 1995; Kim et al., 1995; Ramaniah et al., 1998; Carrillo-Tripp et al., 2003; Lee et al., 2004; Rao et al., 2008; Neela et al., 2012; Biring et al., 2013; Dinh et al., 2014; Soniat et al., 2015; Fifen and Agmon, 2016) studies on hydrated sodium ion clusters have been reported, particularly on the global minima. A global minimum can be determined by obtaining the stabilization energies of isomers. In experiments, comparing enthalpies is the most direct method to obtain thermodynamic information to deduce stabilization energies. With a high-pressure mass spectrometer containing a thermionic alkali ion source, Dzidic and Kebarle reported the enthalpies and entropies of hydrated sodium ion clusters for $n = 1–6$ in gas phase (Dzidic and Kebarle, 1970). With the hydration number increasing, the binding energy per water molecule decreases. Glendening and Feller calculated the stabilization energies and stabilization enthalpies of $Na^+(H_2O)_n$ ($n = 1–6$) at various levels of theory (Glendening and Feller, 1995), in which the RHF and MP2 levels with the 6-31+G* basis set reproduced the experimental values obtained by Dzidic and Kebarle well (Dzidic and Kebarle, 1970).

For the structures of hydrated sodium ion clusters, the coordination number is of particular appeal to studies in different phases. In liquid water, the coordination number of a sodium ion is about 5.5 ± 0.5 based on molecular dynamics simulation (Mancinelli et al., 2007; Megyes et al., 2008; Bankura et al., 2013, 2014; Lev et al., 2013; Galib et al., 2017; Liu et al., 2019). However, in gas-phase clusters, the coordination number is 4 from *ab initio* calculations at 0 K (Kim et al., 1995; Neela et al., 2012; Soniat et al., 2015; Fifen and Agmon, 2016). The structures of $Na^+(H_2O)_n$ ($n = 1–4$), all the water molecules surrounding the sodium ion, were firstly reported by Bauschlicher et al. from *ab initio* calculation (Bauschlicher et al., 1991). For $Na^+(H_2O)_5$, 4+1+0 (the structures of isomers are presented in the form of $n_1+n_2+n_3$, where n_1, n_2, and n_3 are the numbers of water molecules in the first, second, and third solvation shells, respectively) is supported as the global minimum by most *ab initio* calculations at 0 K (Hashimoto and Morokuma, 1994; Kim et al., 1995; Lee et al., 2004; Rao et al., 2008; Neela et al., 2012; Biring et al., 2013; Soniat et al., 2015; Fifen and Agmon, 2016), and 5+0+0 is deemed to be concomitant with 4+1+0 at 298 K (Kim et al., 1995; Fifen and Agmon, 2016). For $n = 6$, at 0 K, several recent *ab initio* calculations stated that 4+2+0 (with D_{2d} symmertry) is the global minimum (Lee et al., 2004; Rao et al., 2008; Biring et al., 2013; Soniat et al., 2015; Fifen and Agmon, 2016), which was proposed by Lybrand and Kollman (1985) based on RWK2 potential (Reimers et al., 1982). However, Neela et al. sustained 5+1+0 to be the global minimum for $n = 6$ calculated at MP2/cc-pVTZ level of theory (Neela et al., 2012). At room temperature, Kim et al. found that 5+1+0 possesses better stability than 4+2+0 by HF/TZ2P (Kim et al., 1995). Differently, Fifen and Agmon indicated that 4+1+1 is dominant, calculated at MP2/6-31++G(d,p) (Fifen and Agmon, 2016).

The infrared (IR) spectra are available in distinguishing the cluster isomers. For hydrated sodium ion clusters, the feature peaks of O–H stretching mode could accurately provide the structure information of clusters (Huang and Miller, 1989; Patwari and Lisy, 2003; Vaden et al., 2004; Miller and Lisy, 2008a,b; Ke et al., 2015), in which the non-covalent interactions, including ion–water interaction and hydrogen bond, weaken the O–H bonds, causing redshifts for O–H stretching vibration modes and producing different feature peaks for different structures (Muller-Dethlefs and Hobza, 2000; Vaden et al., 2002, 2004, 2006; Kozmutza et al., 2003; Bush et al., 2008; Miller and Lisy, 2008a), Using a custom-built, triple-quadrupole mass spectrometer as well as *ab initio* calculations, Lisy et al. reported the IR spectra of $Na^+(H_2O)_n$ ($n = 2–5$) and $Na^+(H_2O)_nAr$ ($n = 2–5$) (Miller and Lisy, 2008a,b; Ke et al., 2015). For $n = 4$, 3+1+0 is the stable structure with bent hydrogen bonds (Miller and Lisy, 2008a). Recently, through straightforward IR spectra for $n = 5$, they speculated that 4+1+0 and 3+1+1 could be concomitant at 75 K (Ke et al., 2015).

In spite of many works having been conducted for $Na^+(H_2O)_n$ ($n = 1–6$), the global minima of $n = 4–6$ remains unclear. Moreover, the non-covalent interaction and electron transfer in hydrated sodium ion clusters have not been discussed in detail, which can elucidate the principle of the shifts of O–H stretching frequency. In this paper, the comprehensive genetic algorithm combined MP2 method is used to determine the global minima of $Na^+(H_2O)_n$ ($n = 1–6$) and simulate their anharmonic vibrational frequencies. Furthermore, charge transfer inside the clusters through NBO analysis and charge density difference are contained, aiming to reveal the principle of non-covalent interactions effecting on the O–H bonds.

METHODS

In this work, some of the structures of the $Na^+(H_2O)_n$ ($n = 1–6$) clusters are adopted from previous literatures (Bauschlicher et al., 1991; Glendening and Feller, 1995; Ke et al., 2015; Fifen and Agmon, 2016). To obtain more isomers for $n = 4–6$, a global search with the comprehensive genetic algorithm (CGA, Zhao et al., 2016) combined with DMol^3 program (Delley, 2000) based on DFT was executed. The CGA method is described in our previous review in detail (Zhao et al., 2016). For each cluster size with $n \geq 4$, we took 10 independent global searches, and for each search, we maintained mating and mutation operations on a population of eight members of up to 3000 GA iterations. Since the Becke-Lee-Yang-Parr (BLYP, Becke, 1988; Lee et al., 1988) functional would provide similar relative energies to MP2 (Møller and Plesset, 1934) method (see **Table S1**), the generalized gradient approximation (GGA) with the BLYP functional and *p*- and *d*- polarization functions (DNP) basis sets were employed to optimize the clusters' isomers in CGA search without symmetry constraint. Considering the calculation cost, zero-point energy (ZPE) was not contained in global search.

Our previous work proved that MP2 is a reasonable method to obtain the energies and properties of small hydrogen-bonded systems (Liu et al., 2013; Shi et al., 2017, 2018). Since the geometrical optimization at augmented correlation-consistent polarized valence double-zeta (aug-cc-pVDZ, Dunning, 1989; Kendall et al., 1992) and aug-cc-pVTZ provide almost the same structures (see **Table S2**), MP2/aug-cc-pVDZ method was utilized to optimize the isomer structures. Single-point energies of these clusters were computed at MP2/aug-cc-pVQZ and MP2/aug-cc-pVDZ levels.

Within harmonic approximation, MP2 calculation usually overestimates the frequencies relative to the experiment, especially for the high frequencies in IR spectra, and may leave out some peaks. Hence, we calculated the IR spectra with anharmonic correction at MP2/aug-cc-pVDZ level at 298 K *via* second-order vibrational perturbation theory (VPT2, Barone, 2005; Barone et al., 2010), as well as to obtain ZPE and thermal correction at 298 K.

For visualizing the bonding strength between the two atoms intuitively, NBO (Carpenter and Weinhold, 1988; Reed et al., 1988) was calculated at MP2/aug-cc-pVQZ level, as well as to obtain the Wiberg bond order (Wiberg, 1968). All the calculations aforementioned were performed in the Gaussian 09 package (Frisch et al., 2013).

Charge density differences of 1+0+0, 2+0+0, 3+0+0, 3+1+0, 4+0+0, and 4+1+0 were calculated using GGA and Perdew–Burke–Ernzerhof (PBE, Perdew et al., 1996) functional, the projector-augmented wave potentials (Blochl, 1994) with an energy cutoff of 500 eV, as implemented in Vienna Ab-initio Simulation Package (VASP, Kresse and Furthmuller, 1996). Only Γ point is k-point with a vacuum layer of over 15 Å was employed in our calculation. The charge density difference is given by:

$$\Delta\rho = \rho_{cluster} - \rho_{Na^+} - \rho_{H_2O} \qquad (1)$$

where $\rho_{cluster}$, ρ_{Na^+}, and ρ_{H_2O} are the charge density of entire hydrate cluster, sodium ion and all the water molecules, respectively.

TABLE 1 | Relative energies (in units of kcal/mol) of $Na^+(H_2O)_n$ (n = 1–6) at 0 K and 298 K calculated at MP2/aug-cc-pVQZ, MP2/aug-cc-pVDZ and BLYP/DNP levels, respectively.

	MP2/aug-cc-pVQZ		MP2/aug-cc-pVDZ		BLYP/DNP	
	0 K	298 K	0 K	298 K	0 K	298 K
1+0+0	0	0	0	0	0	0
2+0+0	0	0	0	0	0	0
3+0+0	0	0	0	0	0	0
4+0+0	0	0	0	0	0	0
3+1+0	1.36	0.04	3.30	1.98	3.32	2.00
4+1+0	0	0	0	0	0	0
3+1+1	2.00	1.48	3.61	3.09	2.68	2.16
3+2+0(1)	2.09	1.50	3.18	2.59	2.47	1.88
5+0+0(1)	2.77	2.51	2.53	2.28	4.16	3.90
5+0+0(2)	2.99	2.41	3.52	2.94	4.81	4.22
3+2+0(2)	2.99	2.68	3.95	3.64	3.55	3.24
4+2+0(1)	0	0	0	0	0	0
4+1+1	1.05	1.72	0.73	1.41	0.14	0.81
4+2+0(2)	1.25	1.40	1.15	1.31	1.63	1.79
4+2+0(3)	2.22	0.76	4.58	3.12	4.48	3.02
5+1+0(1)	2.30	2.15	2.22	2.08	3.76	3.61
4+2+0(4)	3.45	3.28	3.57	3.40	3.37	3.20
6+0+0	4.19	3.92	3.93	3.67	6.56	6.30
5+1+0(2)	4.32	4.66	4.45	4.79	5.10	5.43

RESULTS AND DISCUSSION STRUCTURES

We re-optimized all the isomer clusters obtained from CGA search using the MP2/aug-cc-pVDZ method. The optimized structures and symmetries of $Na^+(H_2O)_n$ (n = 1–6) are present in **Figure 1**. Due to the high computational cost, we used the total energies computed at the MP2/aug-cc-pVQZ//MP2/aug-cc-pVDZ+ZPE level to rank the energy order of all the isomers. **Table 1** lists the relative energies at 0 K and 298 K calculated at MP2/aug-cc-pVQZ, MP2/aug-cc-pVDZ, and BLYP/DNP levels, respectively.

For n = 1–3, the global minima, i.e., 1+0+0, 2+0+0, and 3+0+0, all the water molecules surround the sodium ions and locate at equivalent positions without hydrogen bonds, which are similar to those in previous reports (Hashimoto and Morokuma, 1994; Kim et al., 1995; Rao et al., 2008; Neela et al., 2012; Soniat et al., 2015; Fifen and Agmon, 2016).

For n = 4, CGA has located 3+1+0 and 4+0+0 isomers. Among them, 3+1+0 was the better result in all ten CGA global searches, which possesses lower energy than 4+0+0 without ZPE correction (see **Table S1**). In most previous reports by *ab initio* calculations, 4+0+0 was argued to be the most stable of the structures (Hashimoto and Morokuma, 1994; Kim et al., 1995; Ramaniah et al., 1998; Lee et al., 2004; Rao et al., 2008; Kamarchik et al., 2011; Neela et al., 2012; Fifen and Agmon, 2016), whereas Miller and Lisy reported that the IR spectrum of $Na^+(H_2O)_4$ is similar to that of 3+1+0 at 300 K (Miller and Lisy, 2008a). From **Table 1**, at 0 K, our MP2 calculation manifests that 4+0+0 is lower in energy by 1.36 kcal/mol, while 4+0+0 and 3+1+0 possess almost equal energies at 298 K. As shown in **Figure 1**, 4+0+0 has four equivalent water molecules surrounding the sodium ion, while 3+1+0 is evolved by 3+0+0 connecting a water molecule with two coordinated water molecules by hydrogen bonds.

For n = 5, six isomers are contained in our calculation: 4+1+0, 3+1+1, two 3+2+0 structures, and two 5+0+0 structures with different symmetries. 4+1+0 was the best structure in all ten CGA searches, and had the lowest energy at 0 K. 3+1+1, 3+2+0(1), and 5+0+0(1) possess the relative energies of 2.00, 2.09, and 2.77 kcal/mol, respectively. Besides, 3+2+0(2) and 5+0+0(2) have equal relative energies of 2.99 kcal/mol. At 298 K, 5+0+0(2) possesses a lower energy than 5+0+0(1), and the energetic order of the other isomers doesn't change. At the BLYP/DNP level of theory, 3+2+0(1) becomes the second lowest energy structure, and 3+2+0(2) has lower energy than 5+0+0(1) and 5+0+0(2) at 0 K. In **Figure 1**, it is noteworthy that the global minimum structure 4+1+0 has an extra water molecule located at outer shell of 4+0+0. 3+1+1 structure has a water molecule connecting with the outer water molecule in 3+1+0 *via* a hydrogen bond. Similarly, 3+2+0(1), with C_s symmetry, has a water molecule connecting the isolated coordinated water molecule in 3+1+0, and 3+2+0(2) has a water molecule located beside a coordinated water molecule with a hydrogen bond of 3+1+0. 5+0+0(1) structure has three water molecules form a water cycle via hydrogen bonds, and the other two coordinated water molecules are isolated opposite the water cycle. Differently, 5+0+0(2) has only one isolated coordinated water molecule,

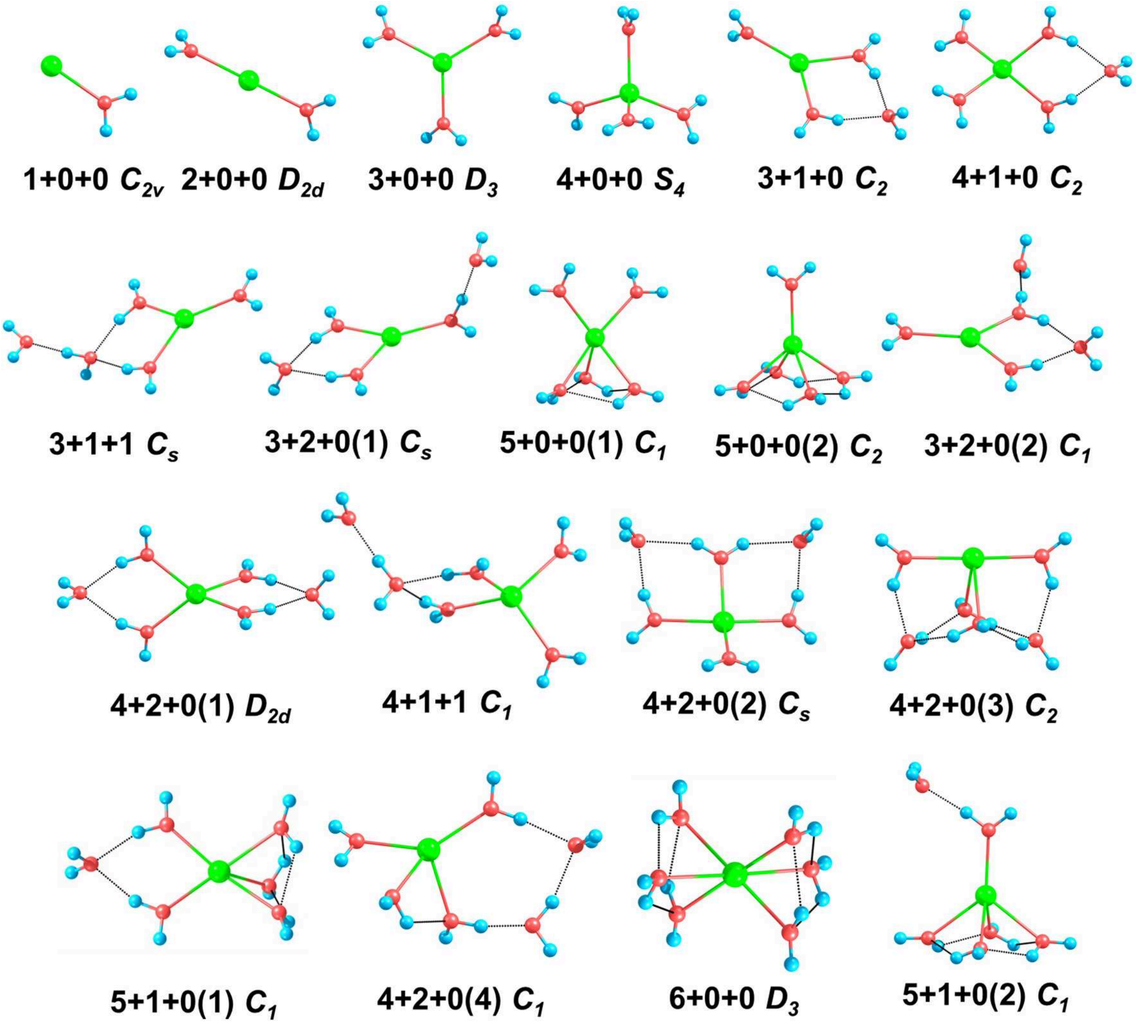


FIGURE 1 | The structures and symmetries of $Na^+(H_2O)_n$ (n = 1–6) optimized at MP2/aug-cc-pVDZ level of theory. Blue, red, and green balls denote hydrogen, oxygen, and sodium atoms, respectively. The black dashed lines represent hydrogen bonds.

with the other four water molecules constituting a quaternary water cycle.

For $n = 6$, we found eight isomers, 4+1+1, 6+0+0, two 5+1+0 structures, and four 4+2+0 structures with different symmetries: D_{2d}, C_s, C_2, and C_1. In all ten CGA searches, 4+2+0(3) had the best solution with the lowest energy at MP2/aug-cc-pVQZ without ZPE (see **Table S1**). From **Table 1**, at 0 K, 4+2+0(1) has the lowest energy, while the relative energies of 4+1+1 and 4+2+0(2) are 1.05 and 1.25 kcal/mol, respectively. The other five isomers possess relative energies of over 2 kcal/mol. At 298 K, 4+2+0(3) becomes the second lowest energy structure rather than 4+1+1, with the relative energy of only 0.76 kcal/mol. Four isomers with four coordinated water molecules have lower stabilization energies both at 0 K and 298 K, indicating that four coordination is more favorable for $n = 6$. Compared to MP2/aug-cc-pVQZ, the calculations at BLYP/DNP shows that 4+2+0(4) has lower energy than 4+2+0(3) and 5+1+0(1). Besides, 6+0+0 possesses the highest relative energy of 6.56 kcal/mol, which is obviously higher than the total energy interval at MP2/aug-cc-pVQZ (4.32 kcal/mol). Combining with the relative energies of $n = 5$, BLYP/DNP gives the same global minima and similar energetic order to MP2/aug-cc-pVQZ. However, BLYP/DNP would overestimate the energies of the 5 and 6 coordinated structures, indicating that the CGA search tends to provide the isomers with 3 and 4 coordinated water molecules. Since the previous *ab initio* calculations show that the coordination number is 4 at 0 K (Kim et al., 1995; Neela et al., 2012; Soniat et al., 2015; Fifen and Agmon, 2016), the CGA search at BLYP/DNP could provide the global minima and other reliable isomers. As seen in **Figure 1**, 4+2+0(1) is a water molecule via hydrogen bonds connecting with the two coordinated water molecules in 4+1+0. Three coordinated water molecules in 4+2+0(2) connect the two outer water molecules *via* hydrogen bonds, and the oxygen atoms in these five water molecules locate in a flat with the sodium ion approximatively. 4+2+0(3) with C_2 symmetry forms a water cycle *via* hydrogen bonds between two coordinated water molecules and the two outer water molecules. Besides, 4+2+0(4) with lowest symmetry has a coordinated water molecule without hydrogen bond. The 4+1+1 structure is a water molecule connecting with the outer water molecule in 4+1+0 *via* a hydrogen bond. In addition, 5+1+0(1) is an extra water molecule located at the outer shell of 5+0+0(1), connecting with the two isolated coordinated water molecules. 5+1+0(2) is just a water molecule connecting with the isolated coordinated water molecule in 5+0+0(2) *via* a hydrogen bond. 6+0+0 could transform from the perfect S_6 symmetry to D_3 symmetry, with two water cycles on two sides of the sodium ion, in accordance with previous calculations based on the polarizable electropole model (Perez et al., 1983).

The bond lengths present interesting variation trends as summarized in **Table 2**. The $\bar{r}$*(Na–O)* increases strictly with the increasing of coordination number, indicating the decreasing of average ion–water interaction. For the structures with two water shells, if a coordinated water molecule is the proton-donor in a hydrogen bond system, the *r(Na–O)* should be shorter. In contrast, if the oxygen atom forms a hydrogen bond, the *r(Na–O)* should become longer. For the *r(O–H)*s, each average *r(O–H)* of water molecules in clusters is longer than the *r(O–H)* of free water molecules (0.966 Å), which stems from the non-covalent ion–water interaction. Meanwhile, the hydrogen bonds also stretch the O–H bonds and make the water molecules asymmetric.

TABLE 2 | Average distances between sodium ions and oxygen atoms [$\bar{r}$*(Na–O)*], distances between sodium ions and each oxygen atoms [*r(Na–O)*] and O–H bond lengths [*r(O–H)*] of coordinated water molecules in $Na^+(H_2O)_n$ (n = 1–6) optimized at MP2/aug-cc-pVDZ level of theory.

	Symmetry	$\bar{r}$(Na–O)/Å	r(Na–O)/Å	r(O–H)/Å
H_2O	C_{2v}			0.966
1+0+0	C_{2v}	2.275	2.275	0.968
2+0+0	D_{2d}	2.302	$2.302^{(2)}$	0.968
3+0+0	D_3	2.335	$2.335^{(3)}$	0.968
3+1+0	C_2	2.322	2.344	0.968
			$2.311^{(2)}$	0.965 0.975
3+1+1	C_s	2.316	2.347	0.967
			$2.300^{(2)}$	0.965 0.979
3+2+0(1)	C_s	2.312	2.287	0.956 0.982
			$2.234^{(2)}$	0.956 0.974
3+2+0(2)	C_1	2.313	2.351	0.967
			2.314	0.965 0.976
			2.273	0.970 0.978
4+0+0	S_4	2.370	$2.370^{(4)}$	0.967
4+1+0	C_2	2.364	$2.378^{(2)}$	0.967
			$2.350^{(2)}$	0.966 0.974
4+1+1	C_1	2.361	$2.384^{(2)}$	0.967
			$2.338^{(2)}$	0.965 0.978
4+2+0(1)	D_{2d}	2.359	$2.359^{(4)}$	0.965 0.974
4+2+0(2)	C_s	2.359	2.379	0.967
			$2.361^{(2)}$	0.965 0.975
			2.337	0.970
4+2+0(3)	C_2	2.411	$2.317^{(2)}$	0.965 0.974
			$2.505^{(2)}$	0.967 0.981
4+2+0(4)	C_1	2.391	2.335	0.968
			2.416	0.966 0.972
			2.339	0.966 0.976
			2.392	0.966 0.984
5+0+0(1)	C_1	2.436	$2.466^{(3)}$	0.967 0.970
			2.408	0.967 0.976
			2.371	0.967
5+0+0(2)	C_2	2.460	$2.486^{(4)}$	0.967 0.972
			2.355	0.967
5+1+0(1)	C_1	2.439	$2.491^{(3)}$	0.967 0.971
			$2.360^{(2)}$	0.965 0.974
5+1+0(2)	C_1	2.472	$2.516^{(4)}$	0.967 0.973
			2.296	0.965 0.981
6+0+0	D_3	2.485	$2.485^{(6)}$	0.967 0.971

The numbers in the parentheses in the line of r(Na–O)/Å represent that there are equivalent water molecules in this number. The two bond lengths in the line r(O–H)/Å represent that hydrogen bond stretches one O–H bond in this water molecule.

CHARGE ANALYSIS

For elucidating the non-covalent interactions in hydrated sodium ion clusters, **Figure 2** and **Table 3** show the NBO overlapping 3D schematic diagrams and electron transfers of 1+0+0. **Figures 2A,B** depict the 2s orbital of sodium ion overlaps the O–H anti-bonding orbitals of water molecule, resulting in electron transfer from the sodium ion to the water molecule. In contrast, **Figures 2C,D** depict the O–H bonding orbitals overlap to the empty orbital of sodium ion, resulting in electron transfer from the water molecule to the sodium ion. From **Table 3**, the amplitude of $E^{(2)}$ manifests that electron transfer from water molecules, including the electrons in O–H bonding orbitals and the oxygen atom's lone pair electron orbitals, to sodium ions is larger than that from sodium ions to water molecules, which synergistically weakens and stretches the O–H bonds of 1+0+0.

For revealing the strength of O–H bonds intuitively, the Wiberg bond order in 1+0+0 (0.740) is smaller than that in a free water molecule (0.790), indicating that the sodium ion weakens the O–H bonds, in accordance with the results from NBO analysis.

To show the charge transfer of the whole clusters directly, the charge density difference of 1+0+0, 2+0+0, 3+0+0, 3+1+0, 4+0+0, and 4+1+0 is presented in **Figure 3**. The electrons from coordinated water molecules assemble at the location between sodium ions and water molecules near the side of oxygen atoms. The electron dissipation mostly happens near the hydrogen atoms, proving that the strengths of O–H bonds become weaker. In addition, the charge density difference of 3+1+0 and 4+1+0 show that the electrons also assemble at the location between the outer water molecules and the sodium ions, indicating that the ion–water interaction also reduces the strength of the O-H bonds in the outer water molecules.

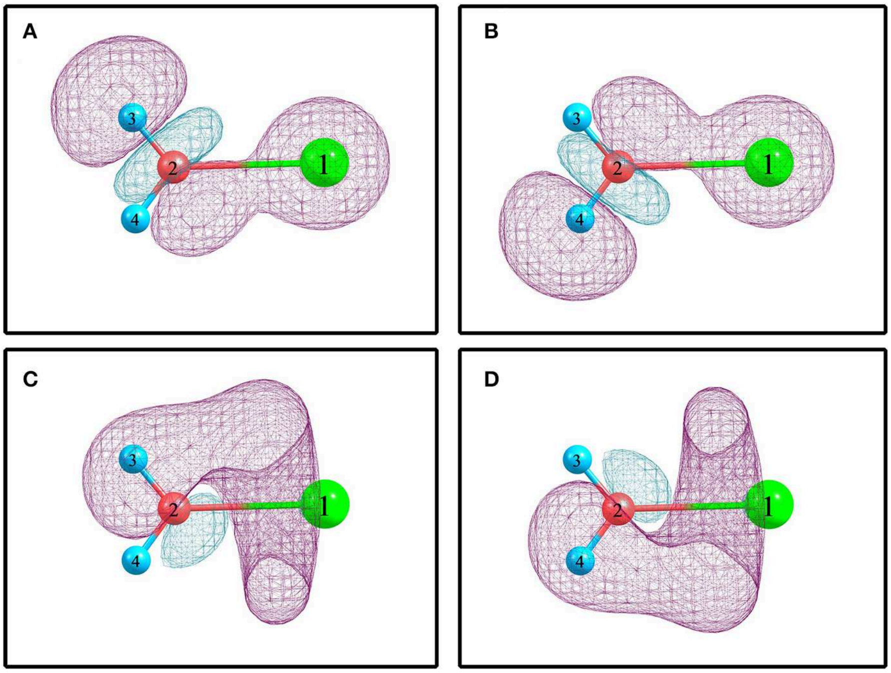


FIGURE 2 | The NBO overlapping and electron transfer in 1+0+0 calculated at the MP2/aug-cc-pVQZ level of theory. **(A)** Na_1^+(s)→ σ^*(O_2-H_3). **(B)** Na_1^+(s)→ σ^*(O_2-H_4). **(C)** σ(O_2-H_3)→ Na_1^+($sp^{0.66}$)*. **(D)** σ(O_2-H_4)→ Na_1^+($sp^{0.66}$)*.

TABLE 3 | The electron transfer and second-order perturbation energies ($E^{(2)}$, expressing the electron delocalization and the extent of charge transfer between different orbitals, in units of kcal/mol) between different natural bond orbitals of 1+0+0 calculated at MP2/aug-cc-pVQZ level of theory.

Donor orbital	Acceptor orbital	$E^{(2)}$
Na_1^+(s)	σ^*(O_2-H_3)	0.09
Na_1^+(s)	σ^*(O_2-H_4)	0.09
Na_1^+(p)	σ^*(O_2-H_3)	0.05
Na_1^+(p)	σ^*(O_2-H_4)	0.05
σ(O_2-H_3)	Na_1^+($sp^{0.66}$)*	0.45
σ(O_2-H_3)	Na_1^+(p)*	0.15
σ(O_2-H_3)	Na_1^+($sp^{1.55}d^{3.85}f^{2.27}g^{1.41}$)*	0.12
σ(O_2-H_4)	Na_1^+($sp^{0.66}$)*	0.45
σ(O_2-H_4)	Na_1^+(p)*	0.15
σ(O_2-H_4)	Na_1^+($sp^{1.55}d^{3.85}f^{2.27}g^{1.41}$)*	0.12
O_2(s)	Na_1^+($sp^{0.66}$)*	0.27
O_2(p)	Na_1^+($pd^{0.35}f^{0.25}g^{0.44}$)	0.20
O_2($sp^{1.02}$)	Na_1^+($sp^{0.66}$)*	1.90

*The * in sodium ion orbitals represent that the orbitals are empty. The subscripts of the atoms correspond to the serial numbers in **Figure 2**.*

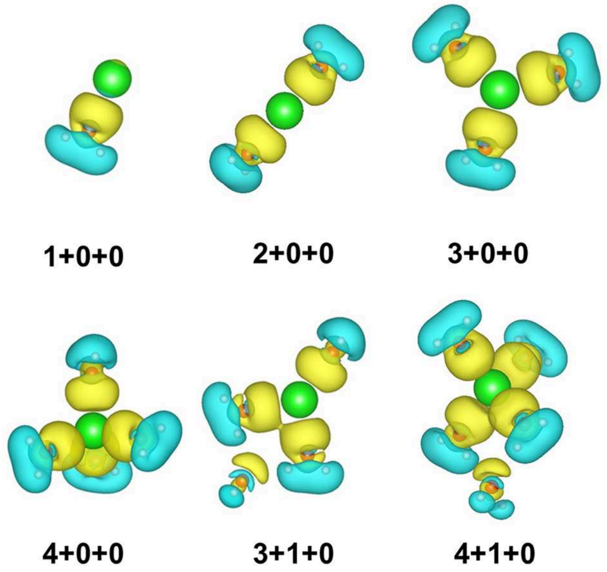


FIGURE 3 | Charge density difference of six small hydrated sodium ion clusters. Yellow and blue spaces represent the electron accumulation and depletion regions, respectively.

VIBRATIONAL SPECTRA

Vibrational spectrum is an intuitionistic method, providing deeper insight into structure differences and non-covalent interactions (Fan et al., 2019), especially O–H stretching vibration modes for hydrated sodium ion clusters. Therefore, the experimental $Na^+(H_2O)_n$ isomers can be determined by comparing the simulated IR spectra and experimental spectra. Our discussion focuses on the high-frequency region (>3,200 cm^{-1}) in IR spectra, which contains the O–H stretching vibration modes and can generally be measured in experiments (Ke et al., 2015).

At first, **Figure 4** shows the IR spectra of 1+0+0, 2+0+0, 3+0+0, and a free water molecule with anharmonic correction. The two O–H stretching vibrational modes are asymmetric (the higher peaks near 3,700 cm^{-1}) and symmetric (the lower peaks near 3,620 cm^{-1}) modes for each structure, and the other peaks are caused by anharmonic correction. Compared to the free water molecule, the asymmetric vibration modes of the three clusters possess redshifts, stemming from the ion–water interactions. The redshifts become smaller with the increasing of coordination number and $\overline{r}(Na\text{–}O)$ in **Table 2**.

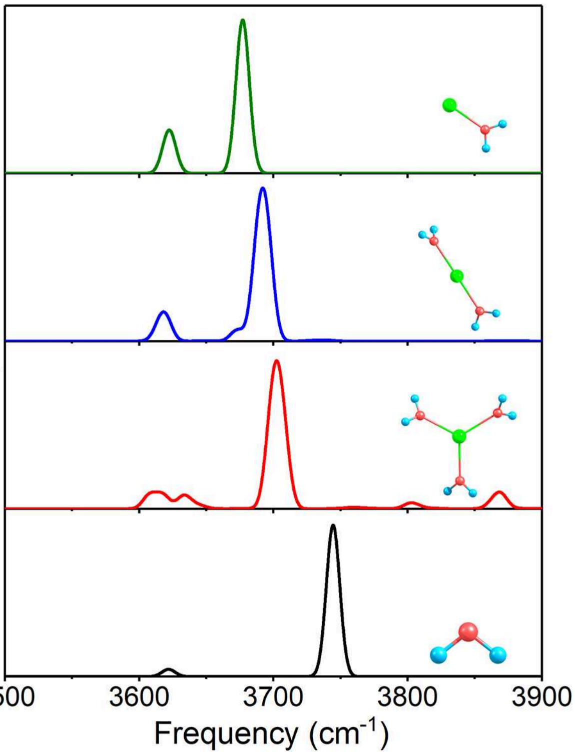


FIGURE 4 | Anharmonic correctional IR spectra of 1+0+0, 2+0+0, 3+0+0 and a free water molecule calculated at MP2/aug-cc-pVDZ level of theory.

For $n = 4$, the spectra of 4+0+0 and 3+1+0, as well as the experimental spectrum of $Na^+(H_2O)_4$ at 300 K (Miller and Lisy, 2008a) are given in **Figure 5**. The two modes of 4+0+0 reproduce the two outstanding peaks of experimental spectrum well. Moreover, the lower peak of the experimental spectrum confirms the small fraction of the existence of 3+1+0. Therefore, 4+0+0 dominates in the experiment at 300 K, and 3+1+0 is concomitant with 4+0+0, in accordance with the almost equal energies at 298 K in **Table 1**.

Figure 6 shows the IR spectra of six isomers for $n = 5$ with anharmonic correction, as well as the experimental spectrum (Miller and Lisy, 2008a). Apparently, no single structure could reproduce the experimental spectrum well. Among, the vibrational modes of 3+1+1 are able to correspond three peaks of the experimental spectrum (Miller and Lisy, 2008a), hence 3+1+1 possesses the most possibility of existing in experiment. However, no mode in 3+1+1 could reproduce the experimental peak near 3,560 cm^{-1}, while all the other five structures have vibrational modes near 3,560 cm^{-1}. Combining with the relative energies in **Figure 1**, 4+1+0, the global minimum, could be the main contributor to the peak at 3,560 cm^{-1}, in accordance with the conclusion in previous reports (Ke et al., 2015; Fifen and Agmon, 2016). Therefore, 4+1+0 and 3+1+1 are concomitant in experiments, which are the two lowest energetic structures at 298 K in **Table 1**.

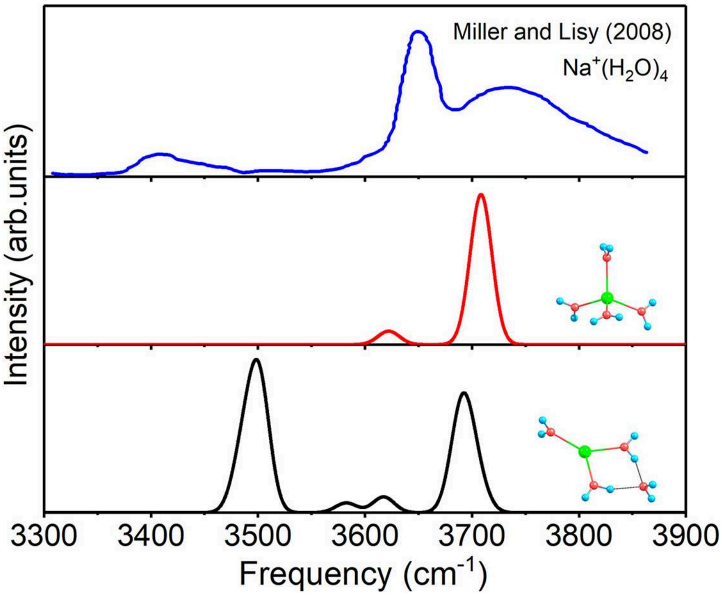


FIGURE 5 | Anharmonic correctional IR spectra of 3+1+0, 4+0+0 calculated at MP2/aug-cc-pVDZ level of theory and experimental spectrum.

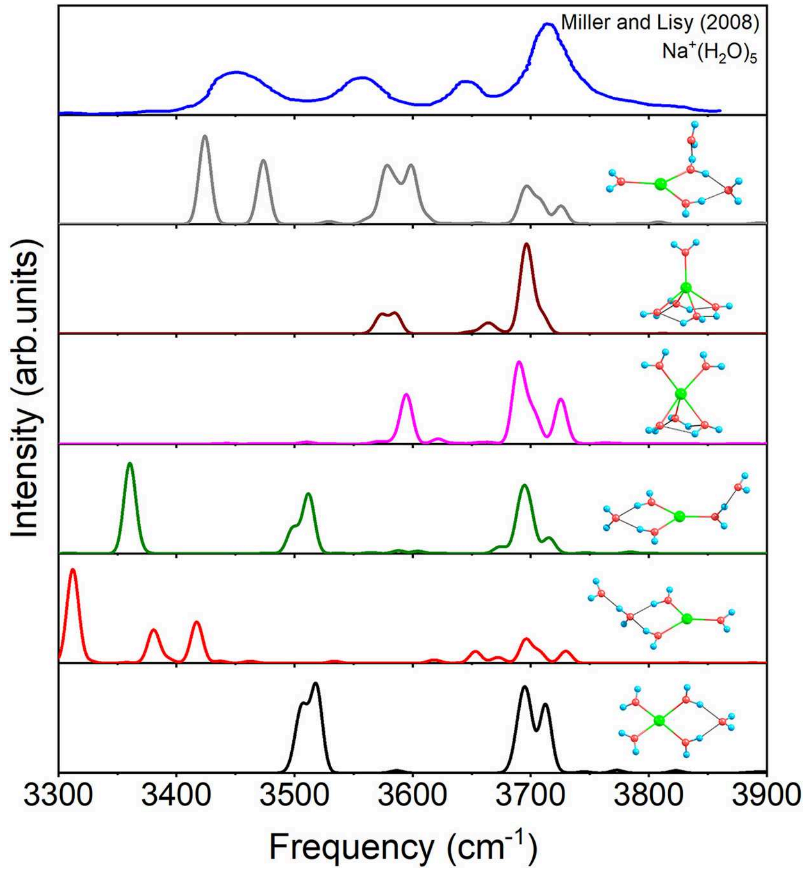


FIGURE 6 | Anharmonic correctional IR spectra of 4+1+0, 3+1+1, 3+2+0(1), 5+0+0(1), 3+2+0(2), and 5+0+0(2) calculated at MP2/aug-cc-pVDZ level of theory and experimental spectrum.

For $n = 6$, **Figure 7** shows the IR spectra of all the eight isomers presented in **Figure 1**. Due to all the coordinated water molecules being equivalent to 4+2+0(1) in **Figure 1**, only two distinct peaks can be observed. Similar to 3+1+1, 4+1+1 possesses an obvious peak at 3340.5 cm^{-1} caused by O–H stretching of the water molecule in the second shell. 4+2+0(2) has no peak under 3,500 cm^{-1} because the hydrogen bonds are not strong enough to make the water molecules distinctly asymmetric, while the three feature peaks, 3371.5, 3395.4, and 3460.2 $cm^{-1,}$ of 4+2+0(3) are generated by the O–H stretching in the water cycle. Because of the hydrogen bonds between the outer molecule and coordinated water molecules in 5+1+0(1), the spectrum has two modes at 3513.3 and 3525.6 cm^{-1}, which can't be found in 5+0+0(1). Due to no equivalent water molecule in 4+2+0(4), the O–H stretching vibration modes with different frequencies make the spectrum more complex than the others. 6+0+0 has a spectrum similar to that of 5+0+0(2) in **Figure 6**, corresponding to the similar water cycles in both structures. Unlike 5+0+0(2), 5+1+0(2) has a significant peak at 3390.0 cm^{-1} which is the O–H stretching mode of the proton-donating coordinated water molecule, indicating that the hydrogen bond between the outer water molecule and the coordinated water molecule is strong in 5+1+0(2).

CONCLUSION

In this work, we investigate the geometries, energies, charges, and anharmonic vibrational properties of $Na^+(H_2O)_n$ (n = 1–6). The CGA search and geometrical optimization for the cluster isomers provide accurate stable structures of $Na^+(H_2O)_n$ (n = 1–6). At 0 K and 298 K, for n = 1–4, all the water molecules in global minima are coordination water molecules, surrounding the central sodium ions. Meanwhile, 4+1+0 and 4+2+0(1) are the global minima of n = 5 and 6, respectively. Thus, the coordination number of global minima of hydrated sodium ion clusters is 4 for $n \geq 4$.

The non-covalent interactions, including ion–water interactions and hydrogen bonds, weaken the O–H bonds, resulting in longer bond lengths, lower bond orders, and redshifts of the O–H stretching mode in IR spectra. The simulated IR spectra with anharmonic correction can reproduce the experimental results well. The results show that 4+0+0 dominates for $Na^+(H_2O)_4$, while 3+1+1 and 4+1+0 should

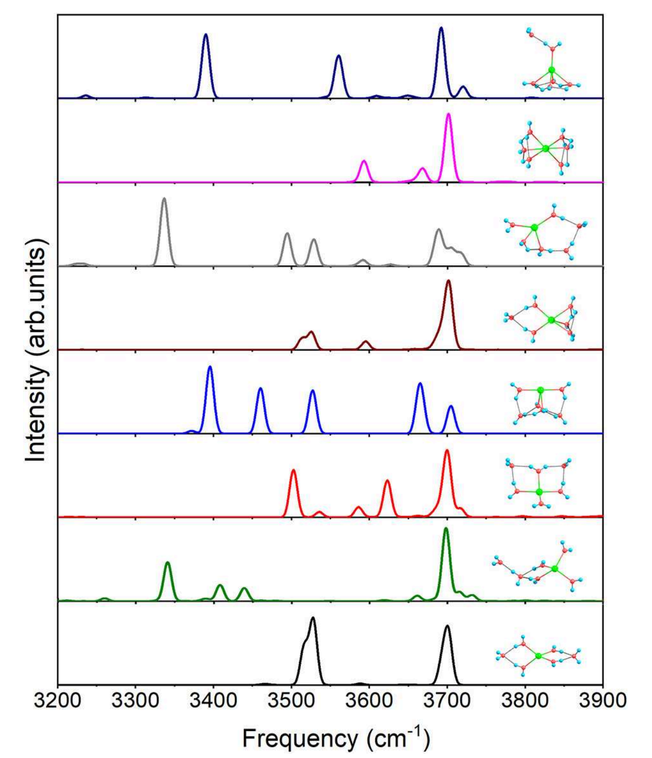


FIGURE 7 | Anharmonic correctional IR spectra of eight isomers for $n = 6$ shown in **Figure 1** calculated at MP2/aug-cc-pVDZ level of theory.

be concomitant for $Na^+(H_2O)_5$ in experiments. The present study executes a believable simulation of structures and vibrational spectra, and provides a comprehensive insight into the non-covalent interactions including ion–water interaction and hydrogen bonds of hydrated sodium ion clusters.

DATA AVAILABILITY

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

PW and RS participated in the design and calculated the data of this study. PW, RS, and YS performed the statistical analysis. LT and XH improved the comprehensive genetic algorithm to adapt the system in this study. YS and JZ carried out the study and collected important background information. All authors have read and approved the content of the manuscript.

FUNDING

This work was supported by the National Magnetic Confinement Fusion Energy Research Project of China (2015GB118001), the National Natural Science Foundation of China (11674046, 11604039), the Science Challenge Project (TZ2016001), the Fundamental Research Funds for the Central Universities of China (DUT16LAB01, DUT16LK16, DUT17GF203, DUT18LK07), and the Supercomputing Center of Dalian University of Technology.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2019.00624/full#supplementary-material

REFERENCES

Arbman, M., Siegbahn, H., Pettersson, L., and Siegbahn, P. (1985). Core electron-binding energies and auger-electron energies of solvated clusters - a computational study. *Mol. Phys.* 54, 1149–1160. doi: 10.1080/00268978500100911

Bankura, A., Carnevale, V., and Klein, M. L. (2013). Hydration structure of salt solutions from ab initio molecular dynamics. *J. Chem. Phys.* 138:014501. doi: 10.1063/1.4772761

Bankura, A., Carnevale, V., and Klein, M. L. (2014). Hydration structure of Na^+ and K^+ fromab initiomolecular dynamics based on modern density functional theory. *Mol. Phys.* 112, 1448–1456. doi: 10.1080/00268976.2014.905721

Barone, V. (2005). Anharmonic vibrational properties by a fully automated second-order perturbative approach. *J. Chem. Phys.* 122:014108. doi: 10.1063/1.1824881

Barone, V., Bloino, J., Guido, C. A., and Lipparini, F. (2010). A fully automated implementation of VPT2 Infrared intensities. *Chem. Phys. Lett.* 496, 157–161. doi: 10.1016/j.cplett.2010.07.012

Bauschlicher, C. W., Langhoff, S. R., Partridge, H., Rice, J. E., and Komornicki, A. (1991). A theoretical study of $Na(H_2O)^+{}_n$ (n=1–4). *J. Chem. Phys.* 95, 5142–5148. doi: 10.1063/1.461682

Becke, A. D. (1988). Density-functional exchange-energy approximation with coorect asymptotic-behavior. *Phys. Rev.* 38, 3098–3100. doi: 10.1103/PhysRevA.38.3098

Biring, S. K., Sharma, R., Misra, R., and Chaudhury, P. (2013). Structural and infrared spectroscopic aspects of ion-water clusters: a study based on a combined stochastic and quantum chemical approach. *J. Cluster Sci.* 24, 715–737. doi: 10.1007/s10876-013-0565-4

Blades, A. T., Jayaweera, P., Ikonomou, M. G., and Kebarle, P. (1990). Studies of alkaline earth and transition metal M^{++} gas phase ion chemistry. *J. Chem. Phys.* 92, 5900–5906. doi: 10.1063/1.458360

Blochl, P. E. (1994). Projector augmented-wave method. *Phys. Rev. B* 50, 17953–17979. doi: 10.1103/PhysRevB.50.17953

Bush, M. F., Saykally, R. J., and Williams, E. R. (2008). Reactivity and infrared spectroscopy of gaseous hydrated trivalent metal ions. *J. Am. Chem. Soc.* 130, 9122–9128. doi: 10.1021/ja801894d

Carpenter, J. E., and Weinhold, F. (1988). Analysis of the geometry of the hydroxymethyl radical by the "different hybrids for different spins" natural bond orbital procedure. *J. Mol. Struc-Theochem.* 46, 41–62. doi: 10.1016/0166-1280(88)80248-3

Carrillo-Tripp, M., Saint-Martin, H., and Ortega-Blake, I. (2003). A comparative study of the hydration of Na^+ and K^+ with refined polarizable model potentials. *J. Chem. Phys.* 118, 7062–7073. doi: 10.1063/1.1559673

Cieplak, P., Lybrand, T. P., and Kollman, P. A. (1987). Calculation of free energy changes in ion–water clusters using nonadditive potentials and the Monte Carlo method. *J. Chem. Phys.* 86, 6393–6403. doi: 10.1063/1.452428

Dang, L. X., Rice, J. E., Caldwell, J., and Kollman, P. A. (1991). Ion solvation in polarizable water-molecular-dynamics simulations. *J. Am. Chem. Soc.* 113, 2481–2486. doi: 10.1021/ja00007a021

Delley, B. (2000). From molecules to solids with the DMol^3 approach. *J. Chem. Phys.* 113, 7756–7764. doi: 10.1063/1.1316015

Dinh, P. M., Gao, C. Z., Klüpfel, P., Reinhard, P. G., Suraud, E., Vincendon, M., et al. (2014). A density functional theory study of $Na(H_2O)_n$: an example of the impact of self-interaction corrections. *Eur. Phys. J.* 68:239. doi: 10.1140/epjd/e2014-40816-1

Dudev, T., and Lim, C. (2010). Factors Governing the Na^+ vs K^+ selectivity in sodium ion channels. *J. Am. Chem. Soc.* 132, 2321–2332. doi: 10.1021/ja909280g

Dunning, T. H. (1989). Gaussian-basis sets for use in correlated molecular calculations.1. the atoms boron through neon and hydrogen. *J. Chem. Phys.* 90, 1007–1023. doi: 10.1063/1.456153

Dzidic, I., and Kebarle, P. (1970). Hydration of the alkali ions in the gas phase. Enthalpies and entropies of reactions $M^+(H_2O)_{n-1}+H_2O = M^+(H_2O)_n$. *J. Phys. Chem.* 74, 1466–1474. doi: 10.1021/j100702a013

Fan, J., Su, Y., Zheng, Z., Zhang, Q., and Zhao, J. (2019). The pressure effects and vibrational properties of energetic material: Hexahydro-1, 3, 5-trinitro-1, 3, 5-triazine (α-RDX). *J. Raman Spectrosc.* 50, 889–898. doi: 10.1002/jrs.5589

Feller, D., Glendening, E. D., Kendall, R. A., and Peterson, K. A. (1994). An extended basis set ab initio study of $Li^+(H_2O)_n$, n=1–6. *J. Chem. Phys.* 100, 4981–4997. doi: 10.1063/1.467217

Fifen, J. J., and Agmon, N. (2016). Structure and spectroscopy of hydrated sodium ions at different temperatures and the cluster stability rules. *J. Chem. Theory Comput.* 12, 1656–1673. doi: 10.1021/acs.jctc.6b00038

Frisch, M., Trucks, G., Schlegel, H., Scuseria, G., Robb, M., Cheeseman, J., et al. (2013). *Gaussian 09, Revision E.01*. Wallingford, CT: Gaussian Inc.

Galib, M., Baer, M. D., Skinner, L. B., Mundy, C. J., Huthwelker, T., Schenter, J. K., et al. (2017). Revisiting the hydration structure of aqueous Na^+. *J. Chem. Phys.* 146:084504. doi: 10.1063/1.4975608

Glendening, E. D., and Feller, D. (1995). Cation-water interactions: the $M^+(H_2O)_n$ clusters for alkali metals M = Li, Na, K, Rb, and Cs. *J. Phys. Chem.* 99, 3060–3067. doi: 10.1021/j100010a015

Hashimoto, K., and Morokuma, K. (1994). Ab initio molecular orbital study of $Na(H_2O)_n$ (n = 1–6) clusters and their ions. Comparison of electronic structure of the "Surface" and "Interior" complexes. *J. Am. Chem. Soc.* 116, 11436–11443. doi: 10.1021/ja00104a024

Hertel, I. I., Huglin, C., Nitsch, C., and Schulz, C. P. (1991). Photoionization of $Na(NH_3)_n$ and $Na(H_2O)_n$ clusters: a step towards the liquid phase? *Phys. Rev. Lett.* 67, 1767–1770. doi: 10.1103/PhysRevLett.67.1767

Huang, Z. S., and Miller, R. E. (1989). High-resolution near-infrared spectroscopy of water dimer. *J. Chem. Phys.* 91, 6613–6631. doi: 10.1063/1.457380

Jensen, F. (1992). Structure and stability of complexes of glycine and glycine methyl analogs with H^+, Li^+, and Na^+. *J. Am. Chem. Soc.* 114, 9533–9537. doi: 10.1021/ja00050a036

Kamarchik, E., Wang, Y., and Bowman, J. M. (2011). Quantum vibrational analysis and infrared spectra of microhydrated sodium ions using an *ab initio* potential. *J. Chem. Phys.* 134:114311. doi: 10.1063/1.3567186

Ke, H., van der Linde, C., and Lisy, J. M. (2015). Insights into the structures of the gas-phase hydrated cations $M^+(H_2O)_nAr$ (M = Li, Na, K, Rb, and Cs; n = 3–5) using infrared photodissociation spectroscopy and thermodynamic analysis. *J. Phys. Chem. A* 119, 2037–2051. doi: 10.1021/jp509694h

Kendall, R. A., Dunning, T. H., and Harrison, R. J. (1992). Electron-affinities of the 1st-row atoms revisited-systematic basis-sets and wave-functions. *J. Chem. Phys.* 96, 6796–6806. doi: 10.1063/1.462569

Kim, J., Lee, S., Cho, S. J., Mhin, B. J., and Kim, K. S. (1995). Structures, energetics, and spectra of aqua-sodium(I): thermodynamic effects and nonadditive interactions. *J. Chem. Phys.* 102, 839–849. doi: 10.1063/1.469199

Kozmutza, C., Varga, I., and Udvardi, L. (2003). Comparison of the extent of hydrogen bonding in H_2O-H_2O and H_2O-CH_4 systems. *J. Mol. Struc-Theochem.* 666, 95–97. doi: 10.1016/j.theochem.2003.08.017

Kresse, G., and Furthmuller, J. (1996). Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* 54, 11169–11186. doi: 10.1103/PhysRevB.54.11169

Lee, C. T., Yang, W. T., and Parr, R. G. (1988). Development of the colle-salvetti correlation-energy formula into a functional of the electron-density. *Phys. Rev. B* 37, 785–789. doi: 10.1103/PhysRevB.37.785

Lee, H. M., Tarakeshwar, P., Park, J., Kolaski, M. R., Yoon, Y. J., Yi, H. B., et al. (2004). Insights into the structures, energetics, and vibrations of monovalent cation-$(Water)_{1-6}$ clusters. *J. Phys. Chem. A* 108, 2949–2958. doi: 10.1021/jp0369241

Lev, B., Roux, B., and Noskov, S. Y. (2013). Relative free energies for hydration of monovalent ions from QM and QM/MM simulations. *J. Chem. Theory Comput.* 9, 4165–4175. doi: 10.1021/ct400296w

Liu, C., Min, F., Liu, L., Chen, J. (2019). Hydration properties of alkali and alkaline earth metal ions in aqueous solution: a molecular dynamics study. *Chem. Phys. Lett.* 727, 31–37. doi: 10.1016/j.cplett.2019.04.045

Liu, Y., Zhao, J., Li, F., and Chen, Z. (2013). Appropriate description of intermolecular interactions in the methane hydrates: an assessment of DFT methods. *J. Comput. Chem.* 34, 121–131. doi: 10.1002/jcc.23112

Lybrand, T. P., and Kollman, P. A. (1985). Water–water and water–ion potential functions including terms for many body effects. *J. Chem. Phys.* 83:2923. doi: 10.1063/1.449246

Mancinelli, R., Botti, A., Bruni, F., Ricci, M. A., and Soper, A. K. (2007). Hydration of sodium, potassium, and chloride ions in solution and the concept of structure maker/breaker. *J. Phys. Chem.* 111, 13570–13577. doi: 10.1021/jp075913v

Mano, I., and Driscoll, M. (1999). DEG ENaC channels: a touchy superfamily that watches its salt. *Bioessays* 21, 568–578. doi: 10.1002/(SICI)1521-1878(199907)21:7<568::AID-BIES5>3.0.CO;2-L

Megyes, T., Balint, S., Grosz, T., Radnai, T., Bako, I., and Sipos, P. (2008). The structure of aqueous sodium hydroxide solutions: a combined solution X-ray diffraction and simulation study. *J. Chem. Phys.* 128:044501. doi: 10.1063/1.2821956

Miller, D. J., and Lisy, J. M. (2008a). Entropic effects on hydrated alkali-metal cations: infrared spectroscopy and ab initio calculations of $M^+(H2O)_{(x=2-5)}$ cluster ions for M = Li, Na, K, and Cs. *J. Am. Chem. Soc.* 130, 15393–15404. doi: 10.1021/ja803666m

Miller, D. J., and Lisy, J. M. (2008b). Hydrated alkali-metal cations: infrared spectroscopy and ab initio calculations of $M^+(H2O)_{(x=2-5)}Ar$ cluster ions for M = Li, Na, K, and Cs. *J. Am. Chem. Soc.* 130, 15381–15392. doi: 10.1021/ja803665q

Møller, C., and Plesset, M. S. (1934). Note on an approximation treatment for many-electron systems. *Phy. Rev.* 46, 618–622. doi: 10.1103/PhysRev.46.618

Muller-Dethlefs, K., and Hobza, P. (2000). Noncovalent interactions: a challenge for experiment and theory. *Chem. Rev.* 100, 143–167. doi: 10.1021/cr9900331

Neela, Y. I., Mahadevi, A. S., and Sastry, G. N. (2012). First principles study and database analyses of structural preferences for sodium ion (Na^+) solvation and coordination. *Struc. Chem.* 24, 67–79. doi: 10.1007/s11224-012-0032-0

Patwari, G. N., and Lisy, J. M. (2003). Mimicking the solvation of aqueous Na^+ in the gas phase. *J. Chem. Phys.* 118, 8555–8558. doi: 10.1063/1.1574018

Payandeh, J., Scheuer, T., Zheng, N., and Catterall, W. A. (2011). The crystal structure of a voltage-gated sodium channel. *Nature* 475, 353–358. doi: 10.1038/nature10238

Perdew, J. P., Burke, K., and Ernzerhof, M. (1996). Generalized gradient approximation made simple. *Phys. Rev. Lett.* 77, 3865–3868. doi: 10.1103/PhysRevLett.77.3865

Perera, L., and Berkowitz, M. L. (1991). Many-body effects in molecular dynamics simulations of $Na+(H_2O)_n$ and $Cl^-(H_2O)_n$ clusters. *J. Chem. Phys.* 95:1954. doi: 10.1063/1.460992

Perez, P., Lee, W. K., and Prohofsky, E. W. (1983). Study of hydration of the Na^+ ion using a polarizable water model. *J. Chem. Phys.* 79:388. doi: 10.1063/1.445534

Pohl, H. R., Wheeler, J. S., and Murray, H. E. (2013). Sodium and potassium in health and disease. *Met. Ions. Life Sci.* 13, 29–47. doi: 10.1007/978-94-007-7500-8_2

Probst, M. M. (1987). A study of the additivity of interactions in cation-water systems. *Chem. Phys. Lett.* 137, 229–233. doi: 10.1016/0009-2614(87)80210-5

Ramaniah, L. M., Bernasconi, M., and Parrinello, M. (1998). Density-functional study of hydration of sodium in water clusters. *J. Chem. Phys.* 109, 6839–6843. doi: 10.1063/1.477250

Rao, J. S., Dinadayalane, T., Leszczynski, J., and Sastry, G. N. (2008). Comprehensive study on the solvation of mono-and divalent metal cations:

Li^+, Na^+, K^+, Be^{2+}, Mg^{2+} and Ca^{2+}. *J. Phys. Chem.* 112, 12944–12953. doi: 10.1021/jp8032325

Reed, A. E., Curtiss, L. A., and Weinhold, F. (1988). Intermolecular interactions from a natural bond orbital, donor-acceptor viewpoint. *Chem. Rev.* 88, 899–926. doi: 10.1021/cr00088a005

Reimers, J. R., Watts, R. O., and Klein, M. L. (1982). Intermolecular potential functions and the properties of water. *Chem. Phys.* 64, 95–114. doi: 10.1016/0301-0104(82)85006-4

Schulz, C. P., Haugstatter, R., Tittes, H. U., and Hertel, I. I. (1986). Free sodium-water clusters. *Phys. Rev. Lett.* 57, 1703–1706. doi: 10.1103/PhysRevLett.57.1703

Schulz, C. P., Haugstatter, R., Tittes, H. U., and Hertel, I. V. (1988). Free sodium-water clusters-photoionisation studies in a pulsed molecular-beam source. *Z. Phys.* 10, 279–290. doi: 10.1007/BF01384862

Shi, R., Huang, X., Su, Y., Si, H.-G., Lo, S.-D., Tang, L., et al. (2017). Which density functional should be used to describe protonated water clusters? *J. Phys. Chem.* 121, 3117–3127. doi: 10.1021/acs.jpca.7b00058

Shi, R. L., Wang, P. J., Tang, L. L., Huang, X. M., Chen, Y. G., Su, Y., et al. (2018). Structures and Spectroscopic Properties of $F^-(H_2O)_n$ with n=1–10 Clusters from a Global Search Based On Density Functional Theory. *J. Phys. Chem.* 122, 3413–3422. doi: 10.1021/acs.jpca.7b08872

Snyder, P. M. (2002). The epithelial Na^+ channel: cell surface insertion and retrieval in Na^+ homeostasis and hypertension. *Endocr. Rev.* 23, 258–275. doi: 10.1210/edrv.23.2.0458

Soniat, M., Rogers, D. M., and Rempe, S. B. (2015). Dispersion- and exchange-corrected density functional theory for sodium ion hydration. *J. Chem. Theory Comput.* 11, 2958–2967. doi: 10.1021/acs.jctc.5b00357

Tang, I. N., and Castleman, A. W. (1972). Mass-spectrometric study of gas-phase hydration of monovalent lead ion. *J. Chem. Phys.* 57, 3638–3644. doi: 10.1063/1.1678820

Vaden, T. D., Forinash, B., and Lisy, J. M. (2002). Rotational structure in the asymmetric OH stretch of $Cs^+(H_2O)Ar$. *J. Chem. Phys.* 117, 4628–4631. doi: 10.1063/1.1503310

Vaden, T. D., Lisy, J. M., Carnegie, P. D., Pillai, E. D., and Duncan, M. A. (2006). Infrared spectroscopy of the $Li^+(H_2O)Ar$ complex: the role of internal energy and its dependence on ion preparation. *Phys. Chem. Chem. Phys.* 8, 3078–3082. doi: 10.1039/b605442k

Vaden, T. D., Weinheimer, C. J., and Lisy, J. M. (2004). Evaporatively cooled $M^+(H_2O)Ar$ cluster ions: infrared spectroscopy and internal energy simulations. *J. Chem. Phys.* 121, 3102–3107. doi: 10.1063/1.1774157

Wiberg, K. B. (1968). Application of pople-santry-segal cndo method to cyclopropylcarbinyl and cyclobutyl cation and to bicyclobutane. *Tetrahedron* 24, 1083–1096. doi: 10.1016/0040-4020(68)88057-3

Zhao, J., Shi, R., Sai, L., Huang, X., and Su, Y. (2016). Comprehensive genetic algorithm for ab initio global optimisation of clusters. *Mol. Simulat.* 42, 809–819. doi: 10.1080/08927022.2015.1121386

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 27 September 2019
doi: 10.3389/fchem.2019.00644

A Global Optimizer for Nanoclusters

*Maya Khatun, Rajat Shubhro Majumdar and Anakuthil Anoop**

Department of Chemistry, Indian Institute of Technology Kharagpur, Kharagpur, India

We have developed an algorithm to automatically build the global minimum and other low-energy minima of nanoclusters. This method is implemented in PyAR (https://github.com/anooplab/pyar) program. The global optimization in PyAR involves two parts, generation of several trial geometries and gradient-based local optimization of the trial geometries. While generating the trial geometries, a Tabu list is used for storing the information of the already used trial geometries to avoid using the similar trial geometries. In this recursive algorithm, an n-sized cluster is built from the geometries of $n-1$ clusters. The overall procedure automatically generates many unique minimum energy geometries of clusters with size from 2 up to n using this evolutionary growth strategy. We have used our strategy on some of the well-studied clusters such as Pd, Pt, Au, and Al homometallic clusters, Ru-Pt and Au-Pt binary clusters, and Ag-Au-Pt ternary cluster. We have analyzed some of the popular parameters to characterize the clusters, such as relative energy, singlet-triplet energy difference, binding energy, second-order energy difference, and mixing energy, and compared with the reported properties.

Keywords: global optimization, PyAR, nanocluster, binary cluster, ternary cluster, nanoalloys, cluster builder

OPEN ACCESS

Edited by:
Emilio Martinez-Nunez,
University of Santiago de Compostela,
Spain

Reviewed by:
Scott Hopkins,
University of Waterloo, Canada
Jadson Claudio Belchior,
Federal University of Minas Gerais,
Brazil

****Correspondence:***
Anakuthil Anoop
anoop@chem.iitkgp.ac.in

Specialty section:
This article was submitted to
Theoretical and Computational
Chemistry,
a section of the journal
Frontiers in Chemistry

***Received:** 31 May 2019*
***Accepted:** 09 September 2019*
***Published:** 27 September 2019*

Citation:
Khatun M, Majumdar RS and
Anoop A (2019) A Global Optimizer for
Nanoclusters. Front. Chem. 7:644.
doi: 10.3389/fchem.2019.00644

1. INTRODUCTION

A major focus in modern nanoscience is to understand the properties of materials on the atomic scale (Eberhardt, 2002). Subnanometer scale metal clusters are of great interest due to their structural and electronic properties (Baletto and Ferrando, 2005), which makes them useful for applications in various field like nanotechnology, electronics, medical device and catalysis (Saha et al., 2012). The atomic clusters may comprise of atoms of the same element such as in fullerenes or atoms of different elements as in nanoalloys (Johnston, 2002). A molecular-level understanding of small nanoclusters would provide insights into the largely empirical field of nanoscience.

Theoretical study of nanoclusters can help us to understand the smooth transition from atoms to bulk materials, especially the size-dependent evolution of the properties (Jortner, 1992; Edwards et al., 1998). The primary input for the theoretical study is their geometry. While determining the geometry of nanoclusters by experiments is extremely difficult, the atomic structure of clusters can be predicted theoretically by geometry optimization tools that are specifically designed for global optimization (Zhao et al., 2017).

Global optimization of functions is an essential part of various research fields and have many real-life applications (Floudas and Gounaris, 2008; Barbati et al., 2012; Khare and Rangnekar, 2013). The global optimization (GO) is the process of finding the best solution, "global maximum" or "global minimum" (GM), based on one or more criteria for a mathematically formulated function (Jäger et al., 2018). The global optimization in our context refers to finding the most stable geometry for a particular cluster, that is the lowest energy atomic arrangements on the potential energy surface (PES). The global minima of atomic clusters (Davis et al., 2015; Shayeghi et al., 2015a) are essential as these are often the most likely structure to be formed in the experiment. Thus, finding the global minimum and other low-lying minima on the PES is helpful to interpret the experimental results (Shayeghi et al., 2014, 2015b; Götz et al., 2016).

The efficiency of geometry optimization (GO) algorithm is crucial for the success in the attempts to understand the cluster science. Some of the popular GO algorithms are Genetic Algorithms (GA) (Johnston, 2003), Basin Hopping (BH) (Wales and Scheraga, 1999), Particle Swarm Optimization (PSO) (Lv et al., 2012; Shi et al., 2019), Artificial Bee Colony (ABC) (Zhang and Dolg, 2015), Simulated Annealing (Kirkpatrick et al., 1983), Threshold Algorithms (Schön et al., 1996) etc. These general GO algorithms are employed in the studies of metal clusters with varying degrees of success. As for any applications of GO, there is no universal method that works for all the molecular systems in chemistry and is an open area of research.

A major challenge in any GO method is the computational complexity, the exponential increase in the search space with system size (Doye and Wales, 1998). A GO algorithm must combine a locally confined search with the wide exploration of the regions without revisiting the same regions (Heiles and Johnston, 2013) in the PES in a computationally effective way. The fine balance of local search and global exploration is required. The re-examination of a minimum only gives redundant information wasting computational resources. On the other hand, confining the search only to a small neighboring area does not allow the algorithm to find the GM in other funnels on the PES. Metadynamics algorithms overcome the revisiting problem by adding time-dependent repulsive bias potential function of collective variables to discourage revisiting the already visited areas. Tabu-search based algorithms (Glover, 1986, 1989, 1990) store the information of previously visited areas to avoid the searching of the already explored region.

In this article, we explain our strategy to find the global minima geometries of atomic clusters—unary, binary and ternary nanoclusters. We have combined two strategies to improve the efficiency: the Tabu-search algorithm to reduce the time spent on the already found minima and a novel recursive approach to reduce the search space by making use of the solutions from the smaller problem. That is, we build the solutions of n sized cluster based on the solutions of $n-1$ sized cluster. This way, the unique geometries of cluster size n can be built bottoms-up starting from the single atom. This method is particularly useful for studying the evolution of structure and properties with the growth of cluster size. We have discussed the implementation and the validation by applying on the known metallic clusters. We have compared the geometries and a few representative properties of the clusters generated by our algorithm with the reported geometries and corresponding properties.

2. THEORETICAL APPROACH

2.1. Cluster Building and Optimization

Our method for the global optimization of the geometries of atomic clusters is an adaptation of our approach for the automated exploration of reaction and aggregation implemented in PyAR (Nandi et al., 2017; Anoop, 2019) program. In this section, we will explain the philosophy and implementation of the `aggregator` modules used for the building of nanoclusters (**Figure 1**). The global optimization for nanoclusters in PyAR involves two parts, generation of several trial geometries and gradient-based local optimization of these trial geometries. In our algorithm, the search for solutions of n-sized cluster make use of the solutions from the search on the $n-1$ sized clusters. At each cycle, the problem is reduced to find the best relative orientations between two species. This approach is analogous to finding the solution of the traveling salesman problem with N cities by adding one more city to the solution of the problem with N-1 cities. The overall procedure automatically generates several unique minimum energy geometries of clusters with size up to n using our evolutionary growth strategy.

This process can be imagined as growing the cluster by adding atoms one by one. The method is similar to the cluster-fusion algorithm of Solov'yov et al. (2004). When the second atom is added to the first one, there is only one possible geometry and there is only one variable—the distance between the atoms. The trial geometry for the dimer is generated as follows. The first atom (called as seed) is placed at the origin of the Cartesian coordinate system. For placing the second atom (named monomer), the value for the x-coordinate is generated as a random number between 0 and 1. Then, the value of x is increased in small steps of 0.1 Å until x is larger than the sum of covalent radii of both atoms ($x > (R_a + R_b)$). This way, the second atom is placed in the X-axis at a distance of no close-contact between the atoms.

The third atom could be placed anywhere in the xy-plane at a distance from the existing atoms of the dimer avoiding close-contacts. Here, the dimer is the seed, and the atom is the monomer. The xy-plane (the search space) is divided into four quadrants. The new atoms are placed in each quadrant sequentially. The quadrant is chosen by generating random numbers for x and y coordinates within a suitable range to fit a particular quadrant. The new coordinates created by these random numbers are normalized so that the point is at a unit distance from the origin. As described above, the third atom initially placed at this position is translated away from the origin to avoid any close contacts.

The search space for the addition of the fourth atom and further on is three-dimensional. The 3D space around the trimer (and larger n'mers) is divided into eight octants. The new atoms are placed in random positions at unit distance from the origin in each of these octants sequentially. The reason for dividing the space into octants is to distribute the new trial geometries evenly so that even with a few trial geometries, there is a chance of exploring different region of space and getting dissimilar geometries. This way, N trial geometries are generated. N is a user-provided parameter. All the trial geometries will be optimized using local, gradient-based optimizers. The optimizations are done by the interfaced software as described later in this section.

Some of the optimized geometries obtained by the gradient-based optimization of these trial geometries may belong to the same minima in the PES, with small differences in geometrical parameters depending on the convergence criteria. Comparison of geometries based on Cartesian coordinates such as RMSD of the atomic positions may fail because the optimization may

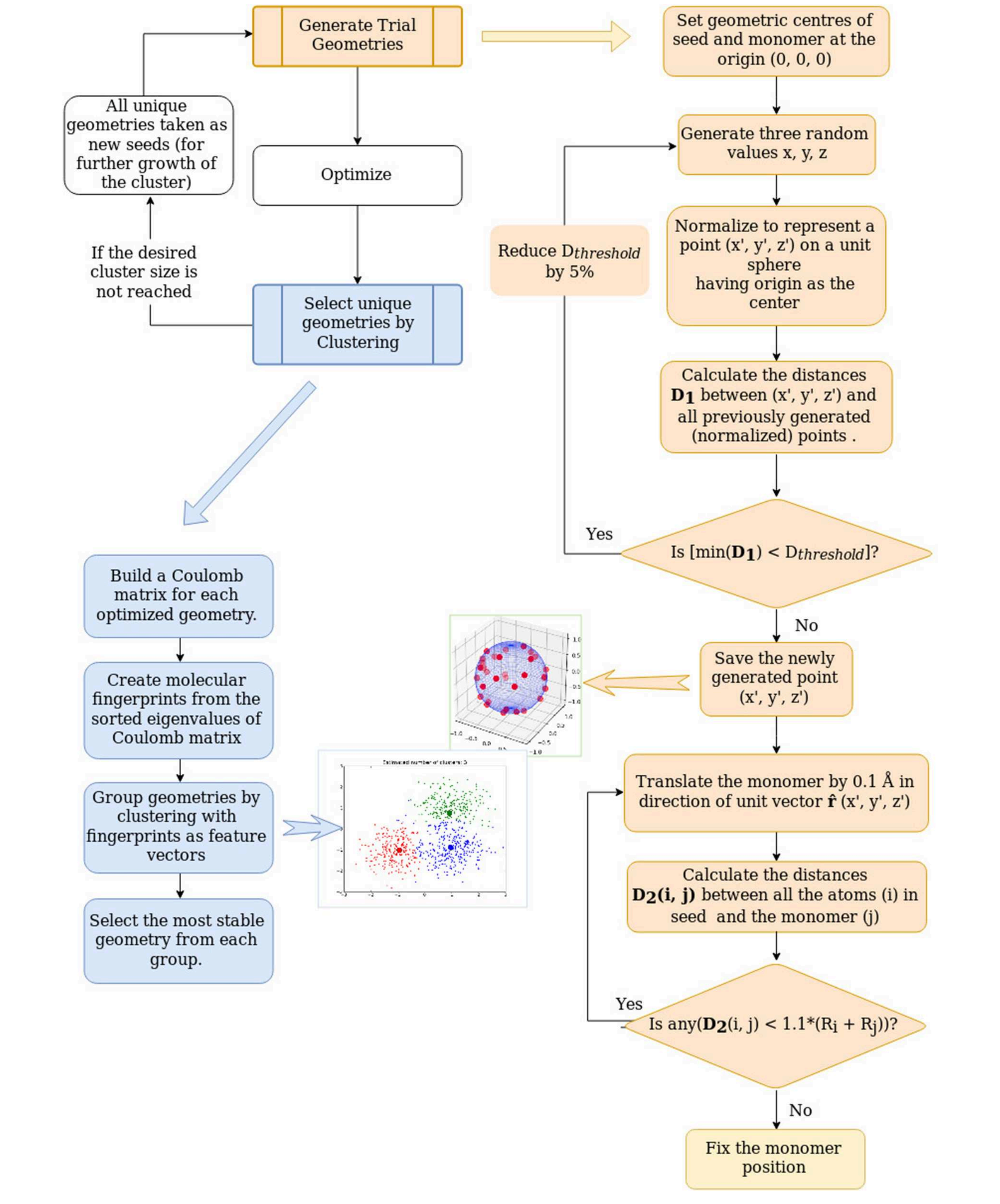

Generate Trial Geometries
Optimize
Select unique geometries by Clustering
All unique geometries taken as new seeds (for further growth of the cluster)
If the desired cluster size is not reached
Build a Coulomb matrix for each optimized geometry.
Create molecular fingerprints from the sorted eigenvalues of Coulomb matrix
Group geometries by clustering with fingerprints as feature vectors
Select the most stable geometry from each group.
Set geometric centres of seed and monomer at the origin (0, 0, 0)
Generate three random values x, y, z
Normalize to represent a point (x', y', z') on a unit sphere having origin as the center
Calculate the distances $\mathbf{D_1}$ between (x', y', z') and all previously generated (normalized) points .
Reduce $D_{threshold}$ by 5%
Yes
Is [min($\mathbf{D_1}$) < $D_{threshold}$]?
No
Save the newly generated point (x', y', z')
Translate the monomer by 0.1 Å in direction of unit vector $\hat{\mathbf{r}}$ (x', y', z')
Calculate the distances $\mathbf{D_2}$(i, j) between all the atoms (i) in seed and the monomer (j)
Yes
Is any($\mathbf{D_2}$(i, j) < 1.1*(R_i + R_j))?
No
Fix the monomer position

FIGURE 1 | The flowchart for the cluster building method.

reorient the molecule, and the Cartesian coordinates are not rotationally invariant. Besides, the same geometry with different ordering of atoms will also be shown as different geometries by such comparisons. Therefore, we have implemented various molecular representations to find the similarity.

One of such representations that we have used in this work is the molecular fingerprints, computed as follows. An n-by-n matrix, known as Coulomb matrix (Rupp et al., 2012; Sadeghi et al., 2013), is made in which the off-diagonal elements are the pairwise Coulomb repulsions $\frac{Z_i Z_j}{R_{ij}}$, and the diagonal elements are $Z_i^{2.4}/2$. The Z_i and Z_j are the core charge of atom i and j. The Coulomb matrix is diagonalized. The sorted eigenvalues are considered as the molecular fingerprints. The fingerprint is used as the feature vector for clustering algorithm (see below) and the euclidean distance between the fingerprints is used as the measure of similarity.

Using the molecular fingerprint representation, these optimized geometries are analyzed and clustered into groups (up to 8 clusters) of similar geometries using clustering algorithms (Nandi et al., 2018) in Scikit-learn (Pedregosa et al., 2011) python library. The most stable geometry from each of these clusters are selected as the minima for this *n*'mer and the most stable among the minima is the global minimum geometry for this *n*'mer. All of these minima are considered for further growth by adding a new atom. This way the degree of freedom of *n*'mer ($3N - 6$) is reduced to 3.

Besides the reduction in complexity, the other significant improvement to increase the efficiency is to avoid revisiting the already visited regions. In our context, we store all the randomly created points and compare the new point with the stored points. For a reasonable comparison, all the positions are generated at a unit distance from the origin, i. e. positions lie on the surface of the sphere of a unit radius (1 Å). If the new position is within the threshold distance from any of the stored positions, the new position is rejected. This threshold distance is initially set as 0.3Å and is increased by 5 % in each cycle. As this idea is adapted from Tabu-search algorithm (Glover, 1986, 1989, 1990), the list of stored positions is referred as the Tabu list. This method of filtering the position makes sure that the trial geometries are sufficiently dissimilar.

The *N* trial geometries created by the method explained above will be optimized with the electronic structure programs that are interfaced with PyAR. Currently we have interfaced with Gaussian 09/16 (Frisch et al., 2016), MOPAC (Stewart, 2016), PSI4 (Turney et al., 2012), ORCA (Neese, 2018), Turbomole (Furche et al., 2014), XTB (Grimme et al., 2017). The user can choose the program and the methods (functional-basis set, semiempirical method). There are few rounds of optimizations. The full set of trial geometries will be initially optimized by loose convergence setting. After filtering similar geometries based on the similarity based on molecular fingerprints, a smaller set of selected geometries will be optimized with standard convergence criteria. In principle, we can also make the automatic procedure to use initial screening with fast and less accurate methods followed by calculations with slow and more accurate methods on a smaller number of geometries.

The methodology described above is for the homometallic clusters. We have extended the procedure to create the binary, ternary and other heteroatomic clusters that are even more interesting and challenging. For making binary clusters, we use both the input atoms as the seed and the monomer instead of one being the seed and the other as the monomer. The procedure, implemented as `binary_aggregator`, generates all combinations of binary clusters of size ranging from A_1B_1 to A_mB_n. The algorithm first treats "A" as the seed and "B" as the monomer and repeats the cycle until the number of "B" atoms reaches *n*. Hence, the row of the matrix is built ranging from A_1B_1 till A_1B_n. When B is considered as seed and A as the monomer, another row is built ranging from A_1B_1 till A_mB_1. Similarly, by using A_xB_y, x < m and y < n, other rows of the matrix can be generated.

We added another layer over the `binary_aggregator` to build the ternary clusters by including a third element. The `ternary_aggregator` operates analogously by adding the element C sequentially to each combination of binary clusters made by `binary_aggregator`. The new monomer is added until it reaches its desired size of the third element. Thus, for each of the binary cluster (A_iB_j; i = 1-m, j=1-n), the 3rd element is added as a monomer to generate ternary clusters ranging from ($A_mB_nC_1$) to ($A_mB_nC_l$) where l is the maximum number of element C.

Current procedures for binary and ternary clusters are expensive because we used exhaustive enumeration. Exhaustive exploration is required until we find some guiding principles for understanding the mixing behaviors of these alloys.

2.2. Properties of Clusters

The relative stabilities of the clusters built using the above described methods can be calculated using the following popular parameters.

2.2.1. Homometallic Clusters

2.2.1.1. Relative energy (RE/eV):

The energy of a cluster compared with the most stable isomer (GM). The higher RE means a lower stability.

2.2.1.2. Singlet triplet energy difference (ΔE_{ST}/eV):

The energy difference between the singlet and triplet state is $\Delta E_{ST} = E_{triplet} - E_{singlet}$. The cluster with a positive ΔE_{ST} has a singlet ground state, and the cluster with a negative ΔE_{ST} has a triplet ground state.

2.2.1.3. Binding energy per atom (BE/eV):

The binding energy per atom (BE or BEPA) is calculated by Equation (1):

$$BE = \frac{1}{N}[E_n - nE_1] \tag{1}$$

where, E_n is energy of n atomic cluster; n is the cluster size or aggregation number; E_1 is the energy of an atom.

2.2.1.4. Second-order energy difference ($\delta^2E(n)$, SOD/eV):
The SOD indicates the higher stability of a cluster of N atoms relative to its heavier and lighter neighbors. Therefore, $\delta^2E(n)$ is more relevant in interpreting experimental mass spectral intensities than the BE (Rogan et al., 2005). Large maxima of $\delta^2E(n)$ shows the higher probability of finding these clusters.

$$\delta^2E(n) = E_{n+1} + E_{n-1} - 2E_n \tag{2}$$

where, E_{n+1} is the total energy of $n + 1$ atomic cluster; E_{n-1} is the total energy of $n - 1$ atomic cluster; E_n is the total energy of n atomic cluster; and n is the cluster size.

2.2.2. Energy Parameters for Binary and Ternary Nanoalloys

2.2.2.1. Binding energy per atom (BE/eV):
The BE for binary and ternary clusters (Song et al., 2005; Demiroglu et al., 2017) is given by Equations (3) and (4):

$$E_b = \frac{1}{N}[E_{tot}(A_mB_n) - mE_{tot}(A_1) - nE_{tot}(B_1)] \tag{3}$$

$$E_b = \frac{1}{N}[E_{tot}(A_mB_nC_l) - mE_{tot}(A_1) - nE_{tot}(B_1) - lE_{tot}(C_1)] \tag{4}$$

where, m, n, and l are the numbers of A, B, and C atoms; $E_{tot}(A_1)$, $E_{tot}(B_1)$, and $E_{tot}(C_1)$ are the electronic energies of a single A, B or C atom and N is the total number of atoms ($N = m + n + l$) in the particular cluster.

2.2.2.2. Mixing energy (ME/eV):
The mixing energy (Song et al., 2005; Pacheco-Contreras et al., 2018) is an indicator of the stability of the binary cluster with respect to its unary counterpart, given by Equation (5):

$$\delta = E_{tot}(A_mB_n) - m\frac{E_{tot}(A_{m+n})}{m+n} - n\frac{E_{tot}(B_{m+n})}{m+n} \tag{5}$$

where, $E_{tot}(A_mB_n)$ is the total energy of the alloy, $E_{tot}(A_{m+n})$ and $E_{tot}(B_{m+n})$ are the total energies of the pure metal clusters, A and B of the same size $(m+n)$. A negative value of δ means a decrease of energy upon mixing and therefore, a favorable mixing.

3. COMPUTATIONAL DETAILS

We used the PyAR program to build the clusters, primarily with the Tight-Binding semi-empirical method, GFN-xTB, with the XTB program (Grimme et al., 2017). This combination is denoted as PyAR|XTB. In a few cases, the selected geometries from PyAR|XTB were reoptimized using PBE0 (Adamo and Barone, 1999) functional and def2-TZVP basis set with the ORCA4.0.1.2 (Neese, 2018) program. These minima from PBE0/def2-TZVP was characterized as true minima with no imaginary frequency. This combined method is denoted as PyAR|XTB||PBE0. We have used another combination where the clusters are built using the ORCA program as the interface using the PBE functional or the BP86 (Perdew, 1986; Becke, 1988) functional and the def2-SVP basis set (Weigend and Ahlrichs, 2005), denoted as PyAR|ORCA. We have added Grimme's dispersion corrections (D3-BJ) (Grimme et al., 2011) in all DFT calculations. We have used effective core potential (ECP) (Pettersson et al., 1983) in the DFT calculations to add the relativistic effect for all the transition metals.

4. RESULTS AND DISCUSSION

We have built various metal clusters—homometallic nanoclusters, bimetallic and trimetallic nanoalloys. In this work, our focus was to validate our approach for its ability to generate the global minimum (GM) and other unique local minima and reproduce the qualitative trends in various properties. Therefore, we have chosen the clusters and alloys that are studied extensively—Pd, Au, Pt, and Al homometallic clusters and Ru-Pt, Au-Pt, Ag-Au-Pt nanoalloys. We have compared the GM geometries and few other low-lying local minima with the corresponding reported geometries. We calculated few properties such as relative energy, binding energy, singlet-triplet energy difference, second-order energy difference, and mixing energy of the clusters and alloys made by our program and compared with the values and trends reported in the literature. Due to the difference in electronic structure theories in different studies, differences are expected in absolute numbers, but overall trends were similar.

4.1. Homometallic Nanoclusters

4.1.1. Palladium

The first example for this study of nanoclusters is the palladium nanoclusters. We have located the unique geometries of Pd_n (n=2–15) clusters using our algorithm implemented in PyAR program. We used two different methods for the global optimization, PyAR|XTB and PyAR|ORCA(PBE). We have also used a two-layer approach in which the search for geometries is done by one method and the selected geometries are optimized again at a different method. For example in the method named as PyAR|XTB||PBE0, the search was done with PyAR|XTB and the geometries selected by this method were further optimized with PBE0. We have employed two more DFT functionals in this study, PyAR|XTB||B3LYP and PyAR|XTB||M06. We have further compared the geometries of Pd_n clusters in singlet and triplet electronic states. The global minimum geometries of singlet Pd_n clusters are shown in **Figure 2**.

Only one minimum was found for triatomic palladium clusters, Pd_3, which has a triangular geometry. The shape of Pd_3 is slightly distorted from the equilateral triangle with the base angle of 59.9°; such non-equilateral geometry was also reported by Nava et al. (2003). The average bond length and bond dissociation energy are 2.54 Å and 2.57 eV at PBE0/def2-TZVP compared to the values from CAS/MRSDCI level calculation (Balasubramanian, 1989) which are 2.67 Å and 3.28 eV.

As the cluster size grew, the program has selected more than one unique structures for clusters with n = 4–15. The relative energies (RE, the energy compared to the global minimum isomer) of all the non-global-minimum geometries are shown in **Figure 2**, along with the results from Nava et al. (2003)

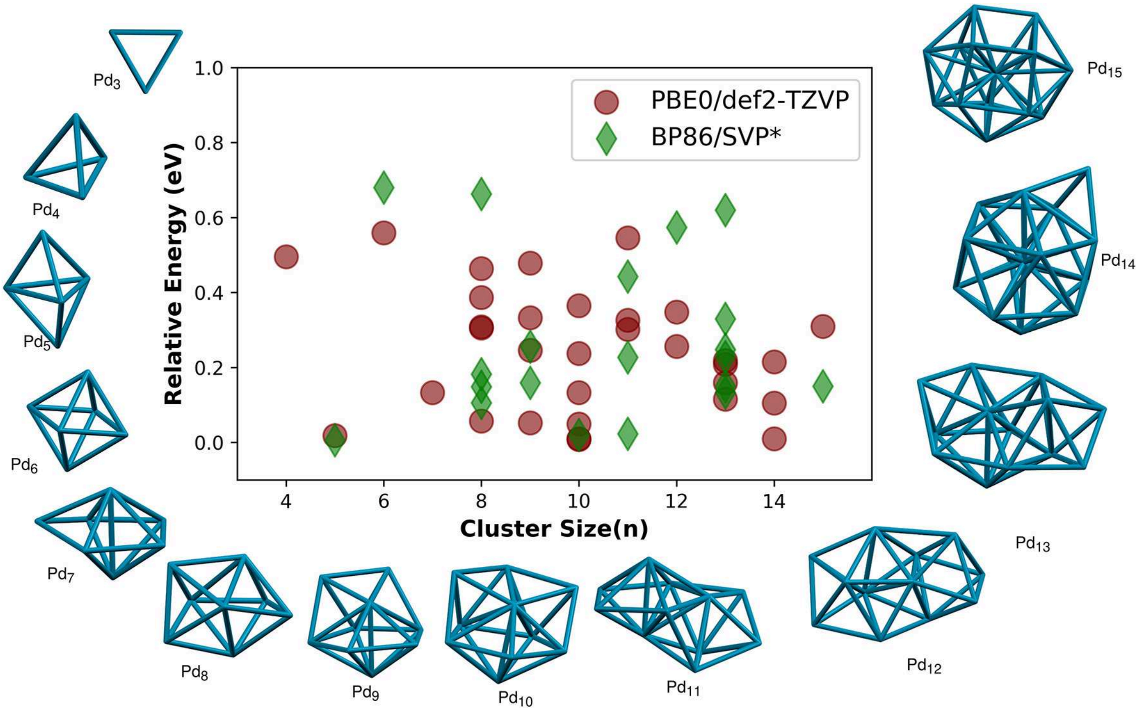


FIGURE 2 | Relative energies (RE/eV), the energy of the optimized isomer compared with the energy of their respective global minimum isomer, of palladium clusters of size ($n = 4 - 15$). The corresponding RE reported using BP86/SVP (Nava et al., 2003) is also plotted for comparison. Global minimum geometries of size $n = 3 - 15$ atoms are shown. The geometries are obtained by using PyAR|XTB calculation followed by optimization at PBE0/def2-TZVP.

for comparison. All the larger Pd_n clusters, $n > 3$, have three-dimensional global minima. Some of these GM geometries are discussed below.

The most stable structure for Pd_4 cluster is tetrahedral. Bond dissociation energy is 4.77 eV at PBE0/def2-TZVP level compared to 5.07eV at the MRSDCI level calculations (Dai and Balasubramanian, 1995). The bond length is 2.62 Å at PBE0/def2-TZVP, 2.68 Å at MRSDCI (Dai and Balasubramanian, 1995) and 2.61 Å using other DFT calculations (Xiao et al., 1999). We found another minimum, a bicyclic, non-planar, *butterfly*-like geometry, not reported before, which is 0.50 eV higher in energy than the tetrahedral GM structure. Global minimum geometry of Pd_5 is trigonal bipyramid. The average bond length in this geometry is 2.74 Å, and the binding energy of the *TBP* structure we calculated at PBE0/def2-TZVP is 1.34 eV, similar to the reported values from the DFT calculation (Wen et al., 2018) using GGA functional (BP/DNP), 2.704 Å and 1.73 eV, respectively. The Pd_6 cluster has an octahedral global minimum. Thus, the most symmetric platonic geometries–trigonal, tetrahedral, trigonal bipyramidal, and octahedral–are the global minima for Pd_3-Pd_6.

The most stable geometry of Pd_7 from the PyAR|XTB calculations is pentagonal bipyramidal (PBP), but is a non-platonic geometry, octahedral core with one cap when PBE and PBE0 methods were used. The PBP was not a minimum, and the trigonal bipyramid with two caps is the next higher energy isomer that has a RE of 0.13 eV compared to GM in PBE0. In the triplet state, the PBP is the most stable structure at BP86 (Nava et al., 2003) and BLYP (Rogan et al., 2005) levels. According to Nava et al. (2003), the mono-capped octahedral and bicapped-TBP Pd_7 are only 0.03eV and 0.05eV higher in energy, respectively, compared to the most stable PBP.

The symmetric dodecahedral geometry was found to be the lowest energy cluster for Pd_8. From Pd_8 to Pd_{13}, pentagonal bipyramidal (PBP) based structures dominate the global minima. For Pd_{13}, the most symmetrical icosahedral structure is not the GM in our calculation (R.E. = 0.21 eV), in agreement with the calculations by Nava et al. (2003) and Reveles et al. (2012) in which the symmetric geometry is higher in energy compared to the most stable geometry by 0.13 eV (BP86/SVP) and 0.16 eV (PBE/DZVP), respectively. The Pd_{14} has an icosahedral core with one cap.

We have calculated the selected geometries in the triplet state as the report (Nava et al., 2003) suggested that many of the Pd clusters have triplet ground states. The ΔE_{ST} is shown in **Figure S1**. In PBE0, all Pd_n clusters have negative ΔE_{ST}, i. e. have triplet ground state, except for Pd atom. The ground state of the Pd atom has a closed-shell electronic configuration. The dimer is well established as a triplet ground state in the literature (Lin et al., 1969; Zacarias et al., 1999; Nava et al., 2003), which is reproduced by our DFT result as well—the singlet Pd_2 has higher energy (0.45 eV) than its triplet state. The dissociation energy of dimer is 0.64 eV which is in agreement with the experimental dissociation energy 0.73 ± 0.26 eV (Lin et al., 1969) as well as

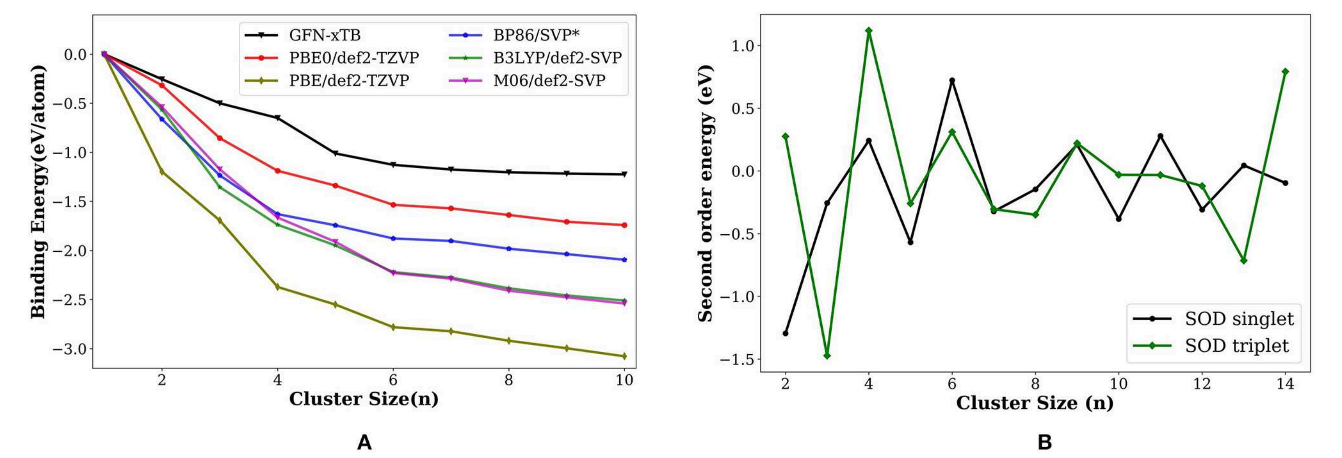


FIGURE 3 | (A) Variation of binding energy (BE; eV/atom) with the cluster size for the most stable palladium cluster obtained with different methods. * values from Nava et al. (2003). **(B)** Second order energy difference (eV) plotted as a function of cluster size (*n*) for the lowest-energy isomers of singlet and triplet state. The geometries obtained using PyAR|XTB calculation were re-optimized at PBE0/def2-TZVP.

various density functional calculations done by Zacarias et al. (1999). The GFN-xTB results, however, showed that all the Pd_n clusters, except Pd_6, have singlet ground state. The ΔE_{ST} in GFN-xTB is large positive for $n = 1, 3$, and 5, but are slightly positive for $n = 2, 4, 7–15$. Thus, ΔE_{ST} is not well represented by GFN-xTB in this Pd_n clusters.

The binding energy per atom (BE/eV) increases as the cluster grows, the trend consistently reproduced by all methods (**Figure 3A**), GFN-xTB, BP86 (Nava et al., 2003), PBE, PBE0, B3LYP, and M06 calculations. The most stable geometries as well as the qualitative features in the overall binding energies gives us a promising strategy for the building of large scale clusters. We can use a two-stage approach where a semiempirical calculations is used for the exploration of minima using PyAR, followed by the optimization in DFT for the selected geometries.

The second order energy difference (SOD; **Figure 3B**) is useful for understanding the stability of cluster with size n compared to the clusters with size $n - 1$ and $n + 1$. The computed SOD for Pd_n cluster shows that Pd_2, Pd_4, Pd_6 are more stable than its neighbors. The clusters with even number of atoms are relatively more stable than the ones with odd number of atoms. This observation is in agreement with Rogan et al. (2005) and Wen et al. (2018) which showed that Pd_2, Pd_4, and Pd_6 are relatively stable than their neighbors.

In short, the study of Pd_n clusters show that the GM structures obtained by our methodology are in good agreement with those from the reported GM structures by other studies (Nava et al., 2003; Rogan et al., 2005). We have studied three more homometallic clusters, Au, Pt and Al, and we have focused different aspects of each clusters below.

4.1.2. Gold

After the study of Pd nanoclusters, we have applied our method to explore the minima of gold clusters using PyAR|XTB(GFN-xTB) and PyAR|ORCA(BP86/def2-SVP). We have generated geometries up to n = 10 with PyAR|DFT and up to 20 with PyAR|GNF-xTB. The GM structures for $n = 4-8$ obtained from our calculations in both the methods are identical with reported structures from CCSD(T) calculations Shi et al. (2010), Baek et al. (2017). Au_4 obtained as a rhombus type structure. The global minimum of gold pentamer is W-shaped, and the hexamer is a planar triangle. The GM of Au_7 has an Au capped the edge of planar triangular Au_6. The Au_8 has GM where an Au is capped to each edges of a square.

For Au_3, the PyAR|BP86 run found triangular and bent geometries. While, the global minima at CCSD(T) level is triangular (Baek et al., 2017), our results at BP86, PBE and PBE0 shows the bent geometry as GM. The bent structure was not a minima with PyAR|GFN-xTB and M06 functional, the optimization resulted in a triangular geometry. Thus, other than Au_3, all the other geometries for Au_n; $n = 4 - 8$ have identical geometries in GFn-xTB and DFT.

The bond length of gold dimer is calculated as 2.472 Å by GFN-xTB and 2.543 Å by BP86 which are in good agreement with the experimental value 2.490 Å. One of the important energy parameters, cohesive energy (CE) of Au_2 is 1.117 eV by our BP86 calculation. This is comparable with 1.1481 eV at the CCSD(T) level (Shi et al., 2010) and 1.1524 eV from experiment (Bishea and Morse, 1991). The CE by GFN-xTB, 4.005 eV, is too high. For the gold trimer, the calculated CE is 1.172 eV, the reported results are 1.161 eV (Shi et al., 2010) and 1.255 eV (James et al., 1994). Au_4 has a CE of 1.487 eV, comparable with the CCSD(T) value of 1.556 eV (Shi et al., 2010). While the results from our BP86 calculations follow the trend with the reported CCSD(T) (Shi et al., 2010) and experimental (Bishea and Morse, 1991; James et al., 1994) results, the GFN-xTB overestimates the CE.

The GM geometries shown in **Figure 4** reveal that the gold clusters have flat GM up to the cluster size of ten atoms. Al_{11} has a 3D geometry. Thus, our approach is able to capture the structure evolution from 2D geometry to 3D geometry that can be attributed to the use of multiple unique seed geometries to build the clusters rather than using only the GM geometry. All the

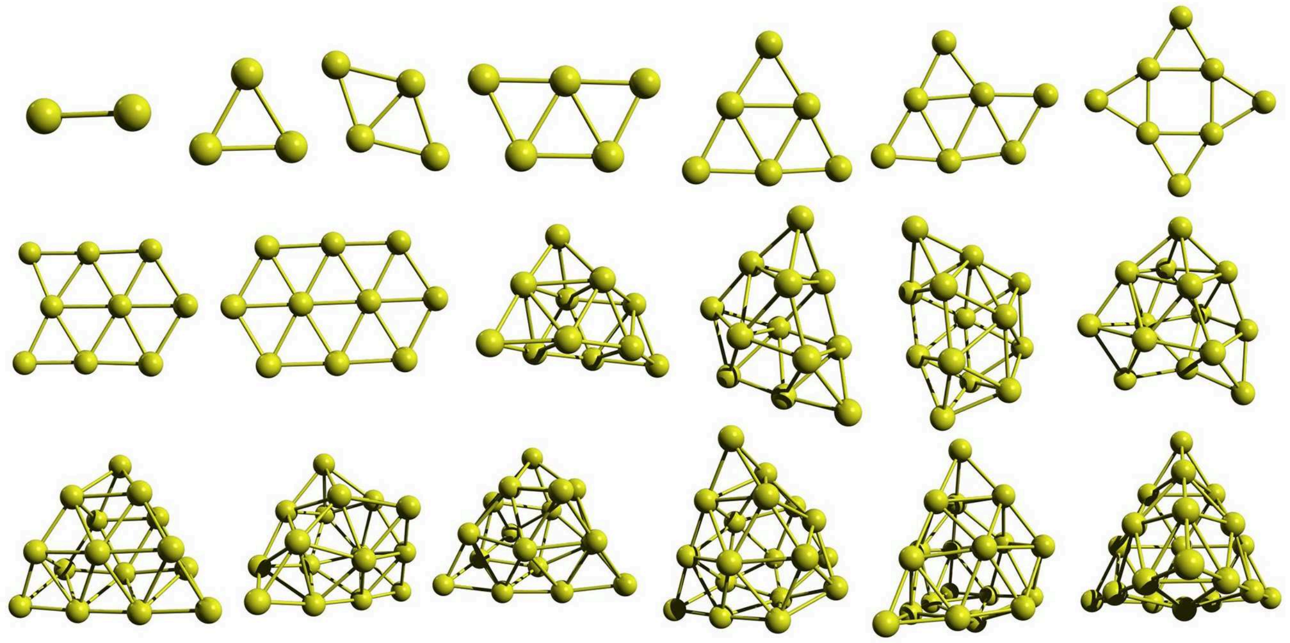


FIGURE 4 | The global minimum structures of Au_n; $n = 2$–20, obtained by the global search using PyAR|XTB.

selected geometries of Au_{10} and Au_{20} is shown in **Figure S2**. The lowest-energy isomers of Au_{10} below 0.4 eV include planar and 3D geometries—the best two are planar. As we have seen above, while the Pd clusters prefer 3D geometries throughout the size range we have studied, the gold clusters remain flat for small sizes, up to 10 in GFN-xTB and BP86 levels.

To study the effect of the number of orientations (**N**) used in the run, we carried out separate runs with different values of **N**. As the size of the cluster increases, the **N** becomes more and more important. For example, the GM (shown with ** in **Figure S2A** produced by one of the PyAR|XTB run with ***N*** = 8 is only one of the local minima, not a GM, in the GA-DFT study (Shayeghi et al., 2015a). However, another run with more orientations along with GFN-xTB resulted in the GM from the GA-DFT and other calculations (Gotz et al., 2013; Shayeghi et al., 2015a). Similar run with DFT also produced the latter GM. The effect is more evident in the Au_{20} cluster.

The Au_{20} has a highly symmetric tetrahedral (T_d) geometry which is one of the most often found structures in the experiment (Gruene et al., 2008) and is one of the most stable geometry in various theoretical calculations (Assadollahzadeh and Schwerdtfeger, 2009; Shayeghi et al., 2015a). The lowest-energy Au_{20} isomers in the range below 0.5 eV are shown in **Figure S2B**). Our geometries are comparable to the ones from previously studied GA-DFT, BH-DFT calculations and the experimental result (Gruene et al., 2008; Shayeghi et al., 2015a; Zhao et al., 2017).

The search for global minimum using only eight orientations was able to locate the tetrahedral global minimum geometry of Au_{20}, however, not always. By varying the number of orientations in the search—***N*** = 8, 16, 32, and 64—we checked the probability of getting the global minimum. When the orientation number is 32, GM structure was found in a single run. As one can anticipate, the possible ways in which the new atom can be added to the $(n-1)^{th}$ cluster increases on increasing the cluster size. Therefore, we have to increase the number of orientation With increasing cluster size. We have illustrated this by plotting the binding energy per atom for the runs with number of orientation as 8, 16, 32, and 64 (**Figure 5A**). We have made an option `auto` for the number of orientation N in which N doubles after each cycle starting with eight in the first cycle, then N increases as 16, 32, 64, 128, 256 and up to a maximum of 512.

4.1.3. Platinum

We studied platinum nanoclusters as the next example as the Pt-based nanoclusters are useful materials with applications in various heterogeneous catalysis. Jennings and coworkers performed GA-DFT searches on small-sized (Pt_n, $n = 3$–6) platinum clusters to find their GM structures. The study showed that Pt clusters have non-singlet ground states, and the geometry of GM's can vary for different spin multiplicity (Jennings and Johnston, 2013). Thus, we have performed three different global minimum searches with multiplicities 1, 3, and 5 on Pt_n; n = 3–6 with PyAR|XTB.

We have observed different global minima for different multiplicities (**Figure 6**) for Pt_4 and Pt_5, in agreement with the GA-DFT study. The Pt_3 has the same triangular geometry in singlet and triplet states, and singlet is the ground state. There are two geometries, **4a** and **4b** for Pt_4. The **4b** has the lowest energy in its singlet state. The ground state of **4a** is a triplet, however, it is higher in energy than the

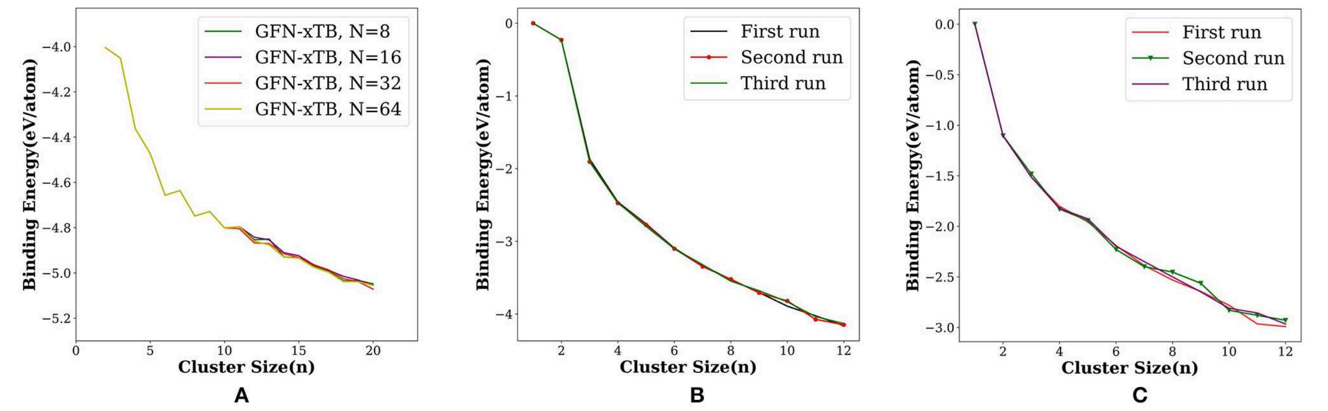


FIGURE 5 | Binding energy per atom (eV/atom) for **(A)** Au_n; n = 2–20 with the number of orientations (N = 8, 16, 32, and 64) and for three different runs done on **(B)** platinum and **(C)** aluminum using PyAR|XTB(GFN-xTB) calculation.

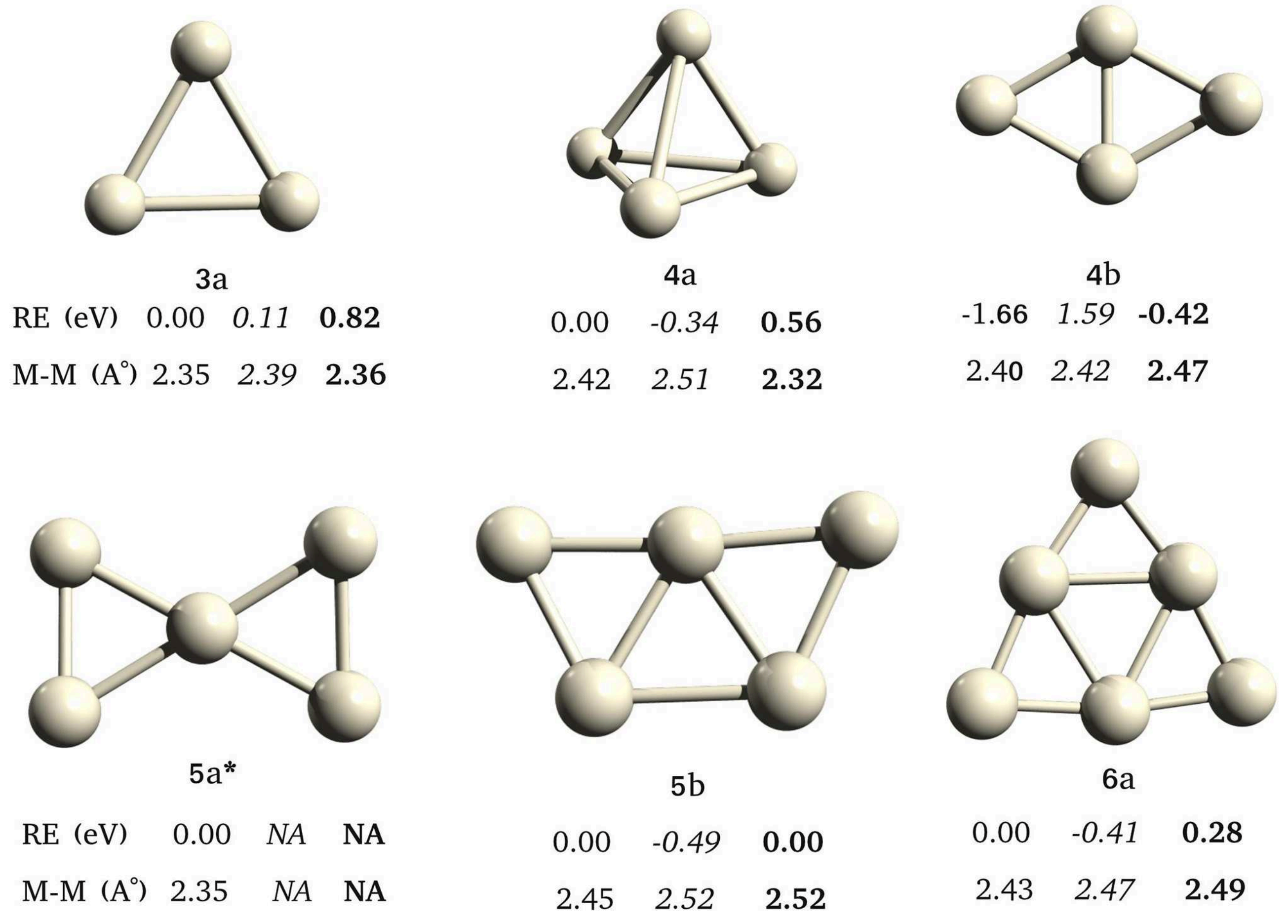


FIGURE 6 | Low energy structures found for pure Pt clusters, from Pt_3 to Pt_6, with different spin multiplicities. *Only singlet state was converged for **5a**. Relative energies (RE/ev) and average bond lengths (Å) of singlet, *triplet*, and **quintet** states shown in normal, italics, and bold fonts.

singlet-**4b**. The **5a** is minima only in singlet state. The GM for Pt_5 is **5b** in triplet state. The Pt_6 has a triplet ground state (**5b**).

4.1.4. Aluminum

The last example for the homometallic cluster in this article is the aluminum cluster. We have built the global minimum structures

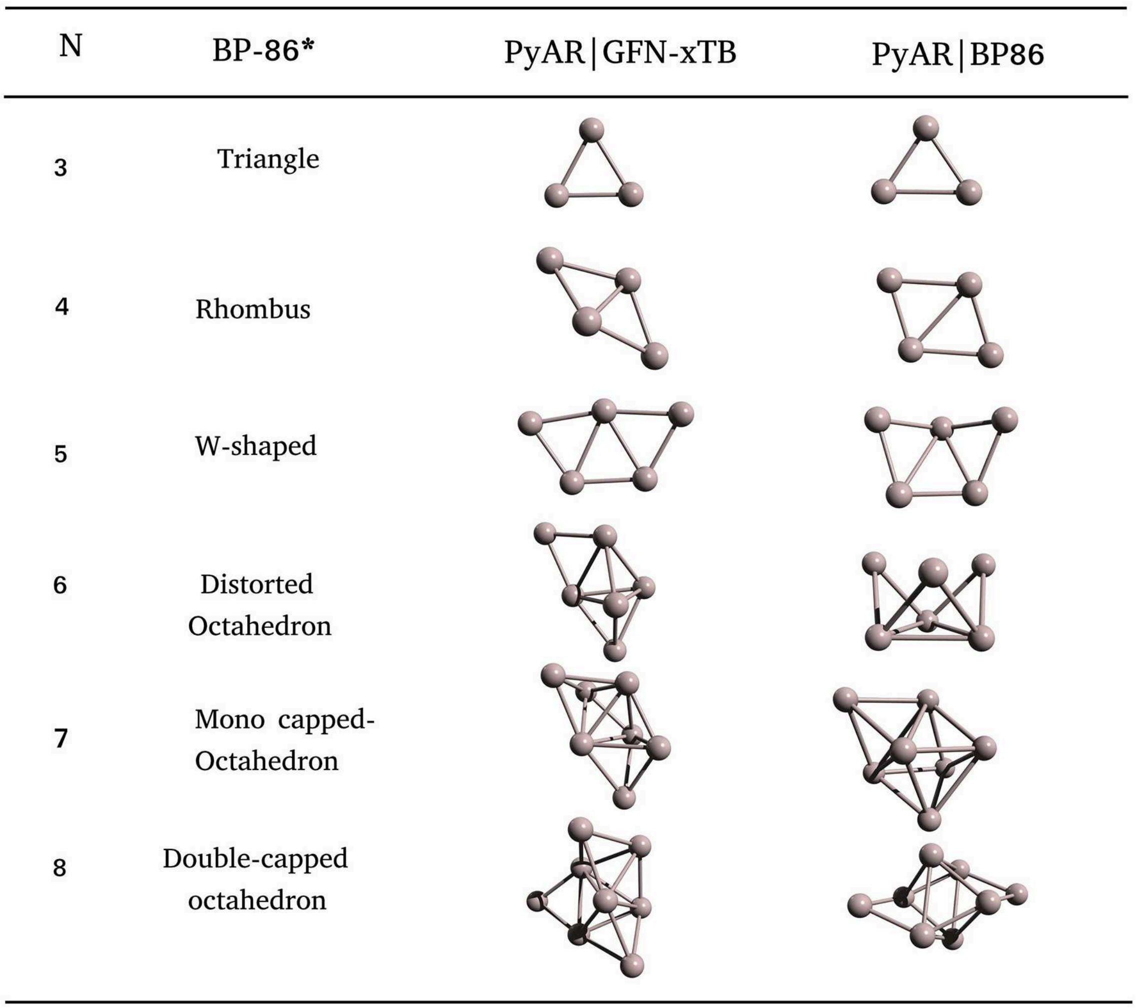


FIGURE 7 | Global minimum structures of Al_n; n = 3–8, obtained by the global search using PyAR|XTB, PyAR|BP86, and reported results (Ahlrichs and Elliott, 1999).

of Al nanoclusters up to Al_{12} with PyAR|XTB, and up to Al_8 with PyAR|BP86. There are several theoretical studies on Al clusters at various levels of theory, such as Sutton-Chen empirical potential (Joswig and Springborg, 2003), DFT (Ahlrichs and Elliott, 1999; Rao and Jena, 1999), and CCSD(T) (Shinde and Shukla, 2014; López-Estrada and Orgaz, 2015).

The most stable structure for the trimer, Al_3 is triangular in both the calculations, PyAR|(XTB, BP86). The bent and linear isomers are 0.53 eV and 0.63 eV higher in energy compared to the most stable structure at BP86/def2-SVP level. We found a planar rhombus geometry for Al_4 with PyAR|BP86 in agreement with the reported *ab initio* methods. PyAR|GFN-xTB calculation gave a slightly different non-planar rhombus geometry as the most stable structure, but tetrahedron is a minima at GA-Sutton Chen potential. The GM of Al_5 is a planar W-shaped structure in our calculation (PyAR|(XTB, BP86)) in agreement with the reported minima from *ab initio* calculations.

The GM of Al_6 by PyAR|XTB is a TBP with an edge-cap, but PyAR|BP86 calculation gave a crown-shaped structure as GM (**Figure 7**). The structure of Al_6 reported by Jones and Ahlrichs (Jones, 1993; Ahlrichs and Elliott, 1999) is a distorted octahedron, not in agreement with any of our minima. For Al_7, the trigonal bipyramid with two capped atoms is the GM at PyAR|XTB, while PyAR|BP86 produced a mono-capped octahedron that matched with the reported minima. The octamer Al_8 showed capped trigonal bipyramid as the minima by PyAR|XTB, octahedral core with two edge-capped by PyAR|BP86 that matched with SC potential (Joswig and Springborg, 2003) and DFT studies (Jones, 1993; Ahlrichs and Elliott, 1999).

4.1.5. General Features

We have studied various features of the approach to finding the global minima of metal nanoclusters. In gold and aluminum clusters, we have compared different methods. All the methods, including semiempirical, produced the same global minima for gold clusters, while the GM was highly dependent on the method for Al clusters. Hence, the choice of appropriate method is crucial.

Some of the clusters have different structural motifs for different sizes. Our method was able to capture the changes in the structural motifs. The global minima for gold clusters were flat upto the size of ten and were 3D geometries afterwards. In order to check these structural changes, we have carried out PyAR|XTB calculation on carbon clusters. We have observed minima corresponding to linear, monocyclic, tricyclic, and the bowl shapes (**Figure S3**).

We have checked the variation in binding energy per atom on varying the number of orientations (**N**) in the Au cluster (**Figure 5A**). The use of more orientations was crucial, especially for the larger clusters. We have then checked the variation in BE for three separate runs for Pt and Al clusters. While the plot of BE for each run (**Figure 5B**) shows nearly perfect overlapping lines for Pt clusters, the BE's sightly differ for Al clusters (**Figure 5C**). As the cluster size increases, the search space increases. Hence, either increase the **N**, or carry out multiple runs, to ensure that most of the local minima are found that increases the chance of finding the global minima. Between these two options, increasing **N** is better as the Tabu list ensures that the trial geometries are dissimilar, while multiple runs may end up in exploring the same local minima more often.

4.2. Binary Clusters

The mixing of two elements may result in properties that are different from the pure forms of each elemental clusters. In the case of binary clusters, we have to consider all different compositions between two elements. Here we have exhaustively explored all combinations in A_iB_j, where $1 \leq i \leq m$; $1 \leq j \leq n$;

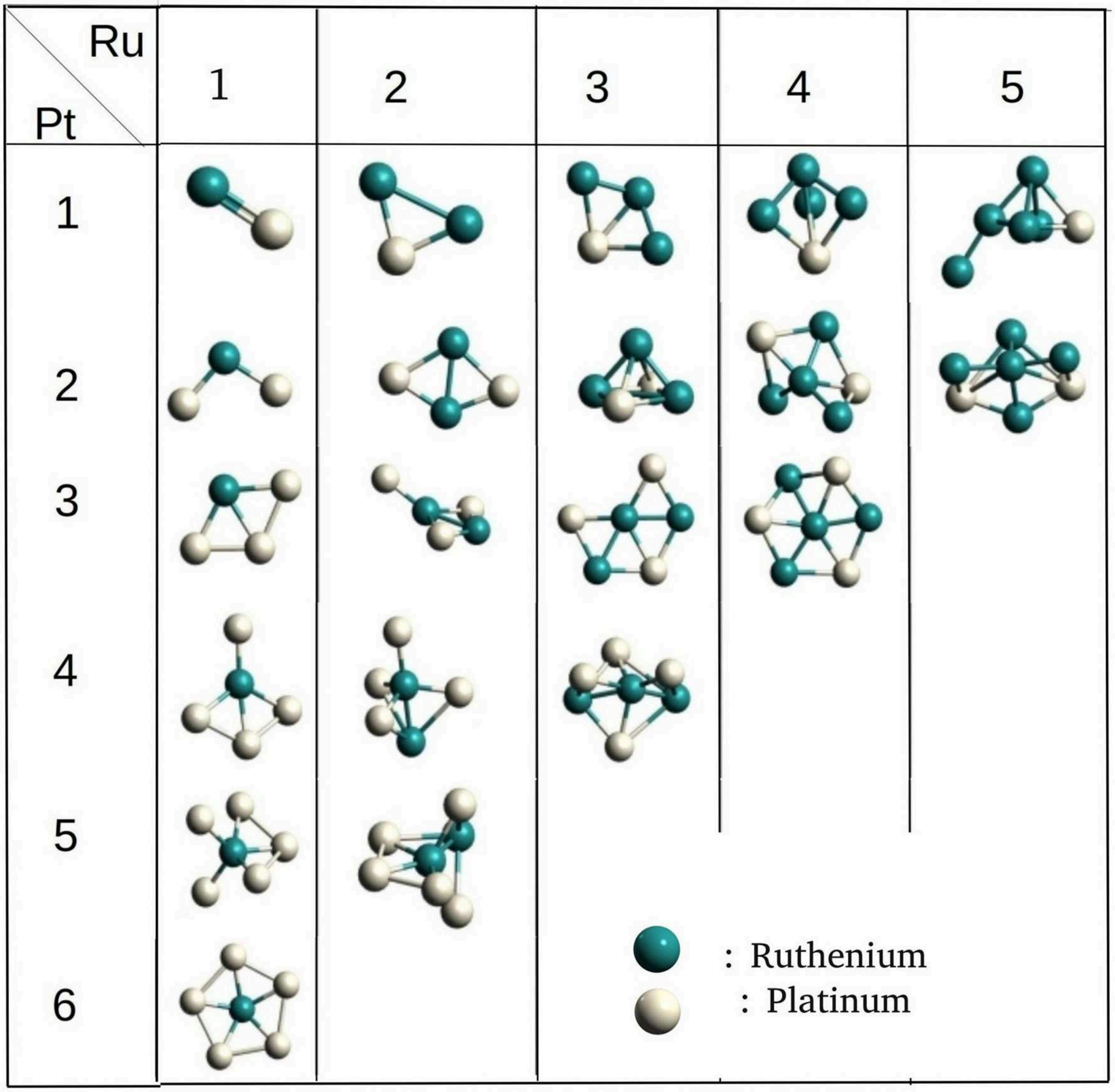


FIGURE 8 | The optimized global minimum geometries of Ru-Pt binary clusters of size 2–7.

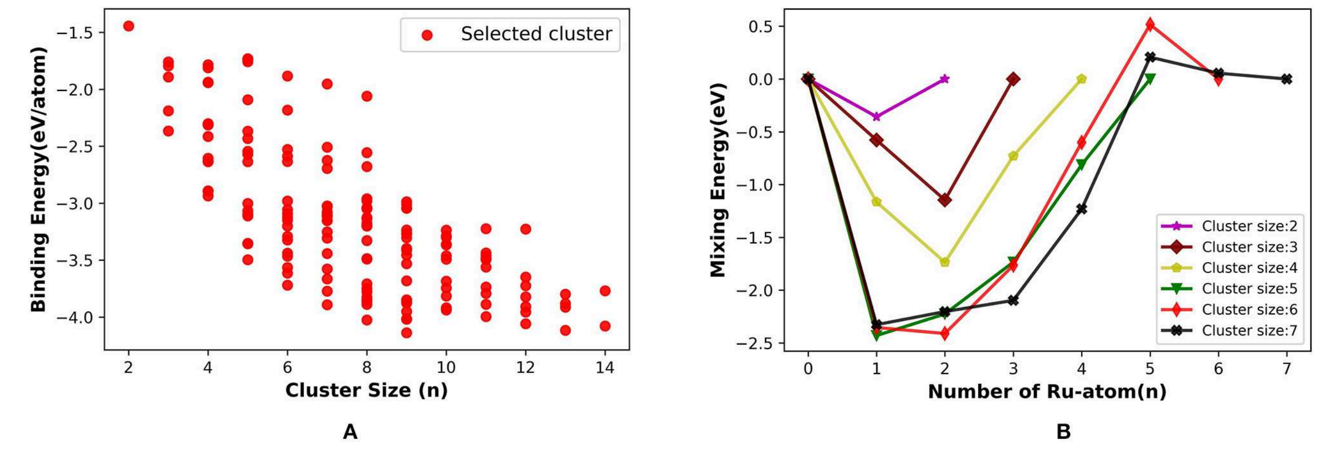


FIGURE 9 | (A) Binding energy per atom (eV/atom) with increasing size from 2 to 14 and **(B)** mixing Energy vs. number of Ru atom for Ru-Pt binary clusters.

the cluster size $N = m + n$, for ruthenium-platinum and gold-platinum binary clusters. One notable feature in the geometries is that, one of the elements tends to become part of the core, while the other tends to be on the surface. The other property of interest is the mixing energy that shows the stability of binary clusters compared to that of the pure unary clusters.

4.2.1. Ruthenium-Platinum Binary Clusters

The binary Ru-Pt nanoalloys showed remarkable enhancement in catalytic activity for CO oxidation (Arico et al., 2001; Liu et al., 2006), compared to when platinum is used in its pure form as a catalyst (Bion et al., 2008), and avoids some of the drawbacks. We applied our method for building binary clusters implemented in `binary_aggregator` in PyAR to build Ru-Pt binary system with the interface to XTB program using GFN-xTB method. We built the Ru-Pt binary clusters up to a total cluster size of 14, i.e., $Ru_1Pt_1 \cdots Ru_7Pt_7$. The lowest energetic clusters are shown in **Figure 8** for a size of 2 to 7.

The general features of the GM geometries match with the reported trend (Demiroglu et al., 2017). In general, the Ru prefers to occupy the core of the clusters with the maximum number of bonds. The Pt, on the other hand, minimizes its number of bonds by seeping on to the surface, having at most three bonds. This observation is in accordance with the higher cohesive energy of Ru (6.74 eV) compared to Pt. (5.84 eV) (Kittel, 2005). The binding energy of the Ru_2 dimer is lower than that of Pt_2, 2.00 eV and 1.94 eV, and the Ru-Pt has the higher binding energy than both (2.13 eV) (Demiroglu et al., 2017). The GM geometries from PyAR|XTB maintain these qualitative features although the individual structures are not identical with the reported structures from Demiroglu et al., as most of these geometries have high spin ground states and we have considered only singlet states (Demiroglu et al., 2017).

For the cluster size of four, all the combinations of $(Ru,Pt)_4$ have similar, non-planar bitriangular geometries. Ru-Ru bond is shorter in Ru_2Pt_2, but two Pt atoms prefer to stay away from each other. For cluster size higher than four, the geometry of GM changes with composition. As the composition of Ru increases in $(Ru, Pt)_5$, the structure changes from planar to 3-D. Similar planar structures were found for Ru_1Pt_4 and Ru_2Pt_3. For cluster sizes with six and seven atoms also, the clusters with a higher composition of Ru have 3-D structures.

The binding energy per atom increases with the cluster size for $(Ru, Pt)_N$ binary cluster in the range that we have considered, up to total cluster size 14. **Figure 9A** shows the average binding energy vs. cluster size of the Ru-Pt clusters, which includes all the selected unique isomers along with GM. The highest BE for each cluster size increases as the cluster grows and gains the highest stability at nine and then again at 13. Our semi-empirical results are in qualitative agreement with the reported DFT results (Demiroglu et al., 2017).

We have calculated the mixing energy (δ), the excess energy of nanoalloy over the pure cluster of the same size, for RuPt binary clusters of size $\mathbf{N} = \mathrm{m} + \mathrm{n} = 2–7$. The effect of mixing Ru with Pt in small clusters calculated as a function of Ru atoms for all compositions of $\mathrm{Ru_mPt_n}$ from $2 \leqslant N \leqslant 7$ clusters are plotted in **Figure 9B**. The mixing is favorable when δ is negative. In our calculation, mixed clusters are more stabilized than the pure clusters except for Ru_5Pt_1 and Ru_5Pt_2. The DFT calculation by Demiroglu et al. (2017), on the other hand, shown positive mixing energy for Ru_3Pt_1. Ru-Pt diatomic molecule is more stable than the pure Ru_2 or Pt_2 dimer. The clusters with one Ru atom (Ru_1Pt_N) more stable than the other possible combinations for $\boldsymbol{N} = 2, 5$, and 7. Two Ru atoms made the binary clusters more feasible when $\boldsymbol{N} = 3, 4$, and 6. Therefore, $(Ru, Pt)_N$ binary clusters with a lesser composition of Ru atoms (one or two) are more favorable in our calculation using semi-empirical method (PyAR|GFN-xTB).

4.2.2. Platinum-Gold Binary Clusters

Platinum-Gold nanoalloys are one of the most studied binary clusters because of their catalytic properties, for example, as a catalyst for CO adsorption (Logsdail et al., 2009; Kaizuka et al., 2010). Song et al. have studied the bonding properties of CO on Pt-Au binary clusters (Song et al., 2005). The catalytic activity of a cluster largely depends on the electronic properties. By

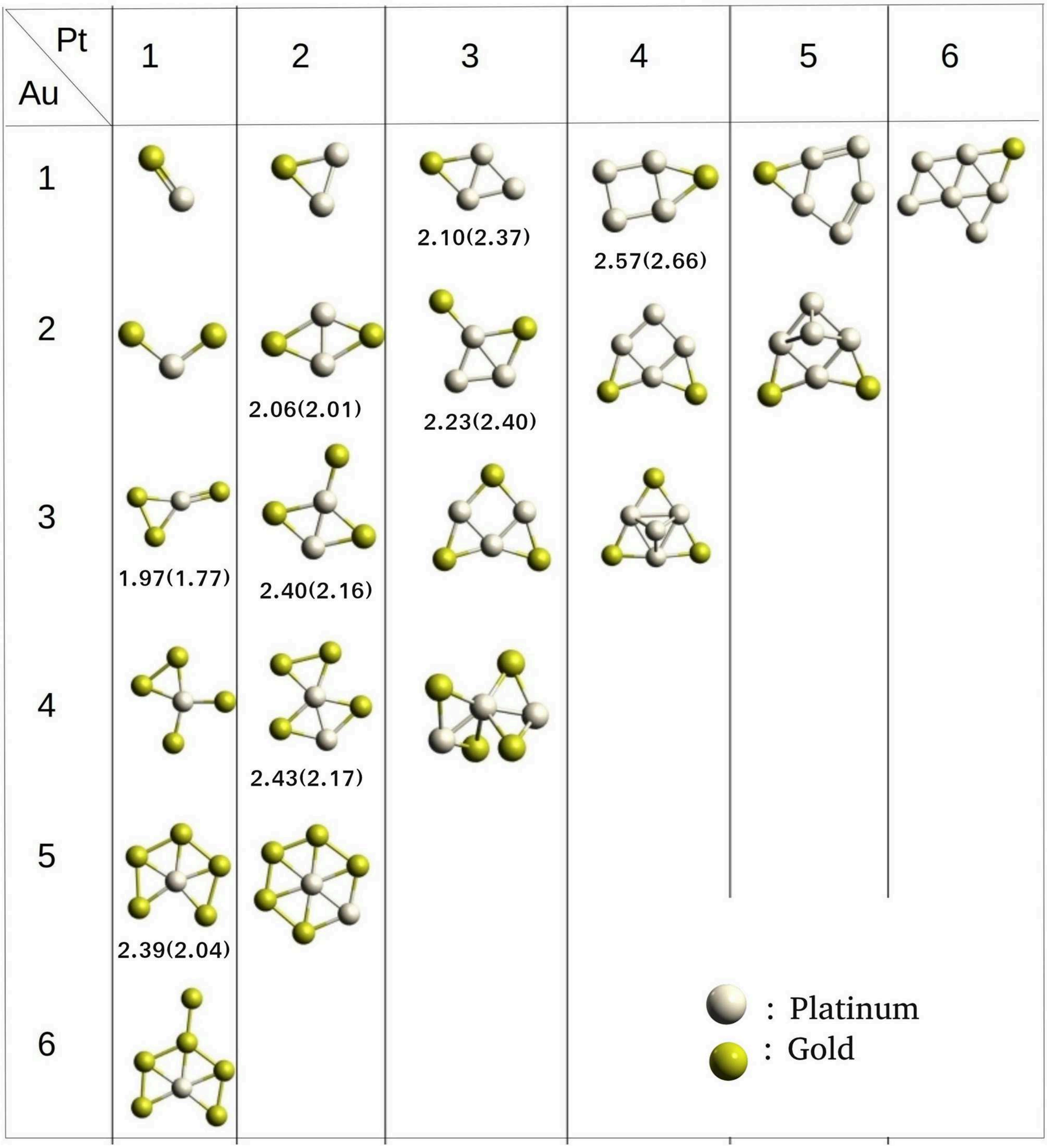


FIGURE 10 | Optimized geometries of Pt-Au clusters from size 2–7 obtained using PyAR|XTB program. Binding energies (eV/atom) from GFN-xTB, and from DFT (PW91/PAW) result (Song et al., 2005) in parathesis.

introducing gold atom in the pure platinum cluster, the electronic properties and thereby, catalytic activity is enhanced.

We have built the (Pt, Au)$_N$ binary clusters; $N = 2$–14 using PyAR|XTB. The lowest energy structures of $N = 2$–7 are shown in **Figure 10**. For (Pt,Au)$_3$ cluster, the Pt$_2$Au has a triangular geometry with Pt-Pt and Pt-Au bonds, while the PtAu$_2$ has a bent structure with both the Au atoms bonded to Pt and has long Au-Au distance. PtAu$_3$ has a planar structure with a triangle of PtAu$_2$ and an exocyclic Au attached to Pt. The other (Pt,Au)$_4$ structures, Pt$_2$Au$_2$ and Pt$_3$Au$_1$ has similar bicyclic

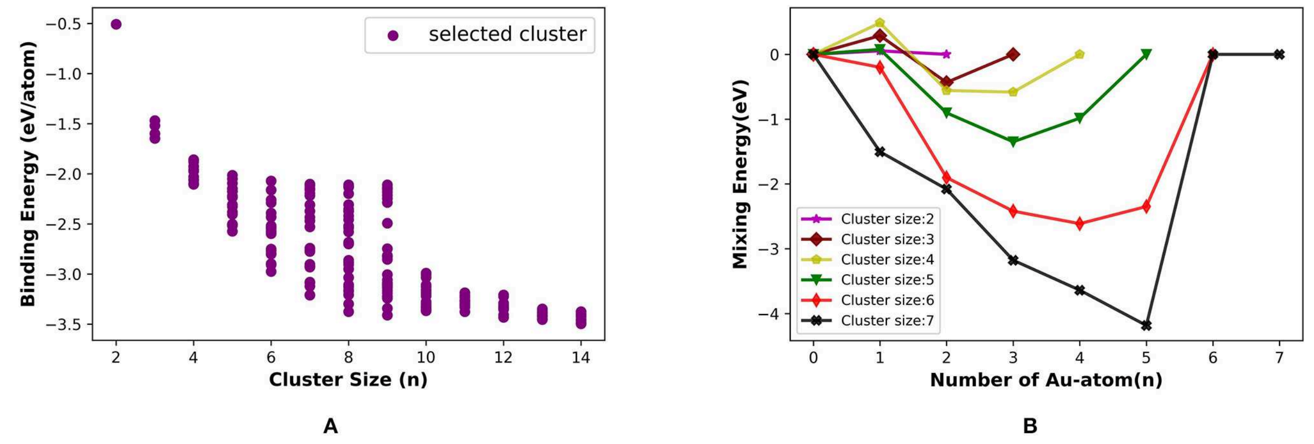


FIGURE 11 | (A) Binding energy per atom (eV/atom) with increasing size from 2 to 14 and **(B)** mixing energy vs. number of Au atom for Pt-Au binary clusters.

quasi-planar structures. Among the (Pt,Au)$_5$ clusters, Au$_2$Pt$_3$and Au$_3$Pt$_2$ where the composition of either gold or platinum is 60% have similar geometries as GM. The (Pt,Au)$_5$ with 80% gold composition makes the structure having a triangular base, but the higher percentage of platinum changes the geometry to a fused four-and-three-membered rings.

As the cluster grows in size, the composition of the alloy will show significant effects on the structure and other properties. For cluster sizes of six and seven atoms, the structures with a higher composition of gold prefer to form planar-like structure. When the composition of Pt is maximum, the cluster tends to acquire a 3D geometry. While Au occupies external sites, Pt occupies the core sites. Apart from these general features, the GM geometries from our study do not match well with the global minimum geometries reported in the literature (Song et al., 2005), due to the different level of theory applied (GFN-xTB vs PW91/PAW) for studying the clusters.

We estimated the average binding energy for (Pt, Au)$_N$ clusters (N = 2–14) using PyAR|GFN-xTB. The cluster gains the highest stability when it reaches the size nine and again at size 14 in **Figure 11A**. We compared our results (PyAR|XTB) with the DFT results by Song et al. (2005). Binding energies of planar Pt$_1$Au$_3$, Pt$_2$Au$_2$, and Pt$_3$Au$_1$ are shown in 10 with the corresponding reported values. The planar minima of Pt$_4$Au$_1$, Pt$_3$Au$_2$, and Pt$_2$Au$_3$ are in agreement with the reported geometries.

We have calculated the mixing energy—the stability of mixed cluster compared to its pure form—for the (Pt, Au)$_N$ clusters. Most of the GM geometries with combinations of Pt$_m$Au$_n$ ($2 \leqslant m + n \leqslant 7$) clusters have negative mixing energy, except for Pt$_1$Au$_2$, Pt$_1$Au$_3$ (**Figure 11B**). Hence, the mixing is, in general, favorable. For cluster size up to seven, the clusters with one or two Pt atoms have the highest stability.

4.3. Ternary Aggregate

The catalytic activity of metal nanoclusters can be enhanced by introducing a second element as well as a third element. Some ternary metal clusters were shown to have higher activity than their unary or binary counterparts (Fang et al., 2011). However, the details of the mechanisms for the enhanced activity is largely unknown as even the structural details of these binary, ternary or other heterometallic clusters are unknown. Finding the global minima of the ternary cluster is even more difficult as compared to the unary and binary systems. Ternary cluster, A$_l$B$_m$C$_n$ (l + m + n = N), can have geometries different from their unary or binary counterparts and can have different structures for different compositions. This high level of complexity in the PES is a challenge for the high-level theoretical calculations to explore the surfaces efficiently. There is a lot to be learnt about the ternary clusters; computational chemistry can serve greatly in this endeavor.

We have studied Platinum-Gold-Silver clusters as an example for a ternary system to validate the `ternary_aggregator` module implemented in PyAR program. We have built the Pt-Ag-Au cluster of total size up to 15 using xTB interface. In the GM geometries, Pt and Ag are near the core, while Au atoms are in the periphery. As we have discussed in the binary systems, these preferences can be attributed to the bond strengths. The bond strength calculated at GFN-xTB level follows the order: Ag-Ag (-5.19 eV) > Pt-Pt (-4.4 eV)> Au-Au(-3.9 eV); the bond energy is given in parenthesis. The geometries of the most stable ternary clusters are shown in **Figure 12**. Most of the structures are quasi-planar or three-dimensional. The general features of the minima are in accordance with the studies by Pacheco-Contreras et al. using Basin Hopping global search with Gupta Potential (as the force field), and using DFT for final optimization for more accuracy (Pacheco-Contreras et al., 2018).

5. CONCLUSIONS

We have developed a methodology to build the unique geometries of nanoclusters and nanoalloys. The clusters are built by adding atoms one-by-one starting from one atom up to the desired size. The following steps are involved in the method: (a) For adding an atom to the N-sized cluster,

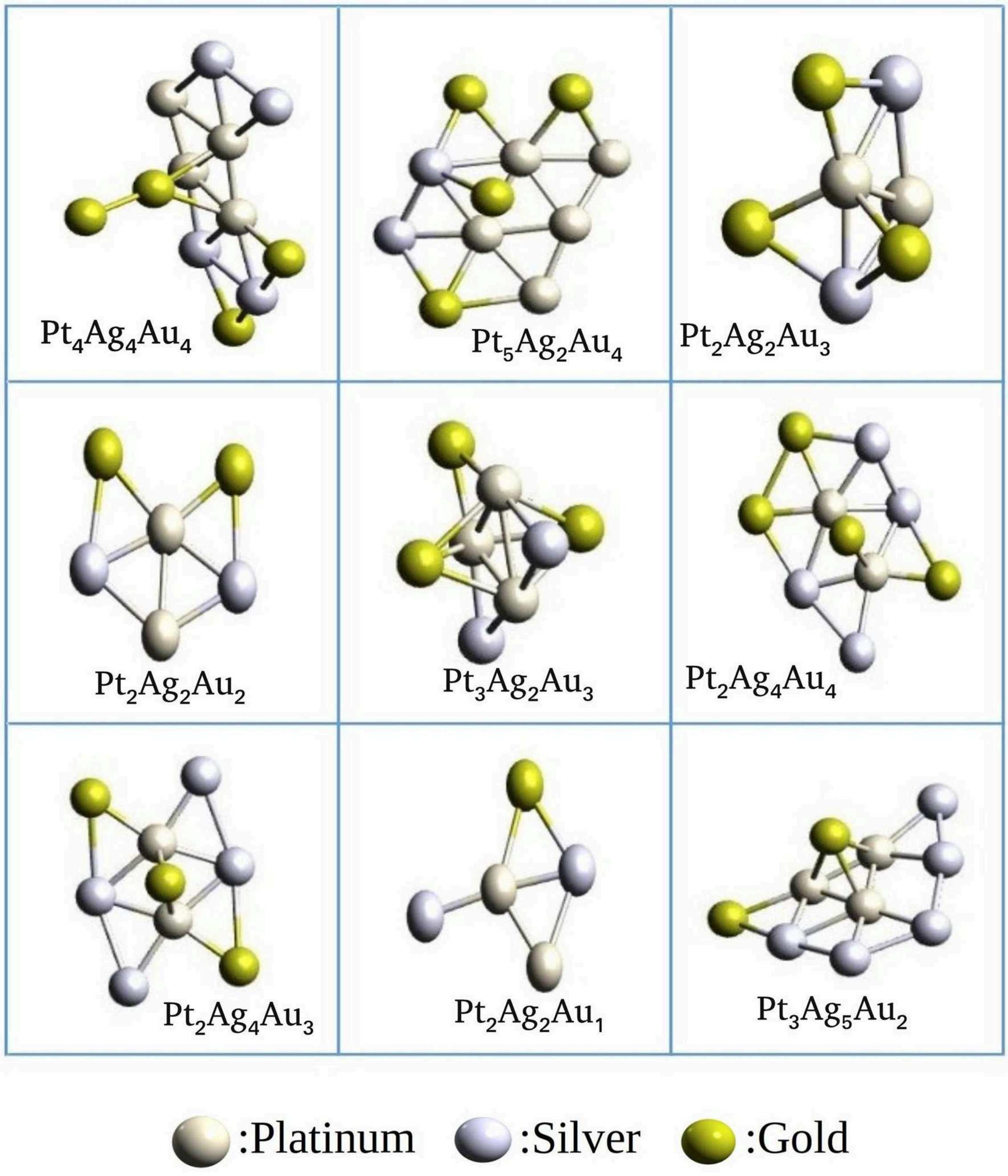


FIGURE 12 | Ternary clusters of Ag, Au, and Pt generated using PyAR|XTB program.

several trial orientations are generated by placing the atom at different random positions around the cluster, (b) These orientations are then optimized by gradient-based methods by the interfaced electronic structure programs, (c) From all the optimized geometries, the similar geometries are removed, the unique structures are selected by clustering algorithms, and these selected geometries are used for the next cycle. These steps (a–c) are repeated to add an atom to all the selected seed molecules. This atoms are added until the cluster grows to the desired size. The similarity between the molecules are compared using the molecular representation based on fingerprints of the Coulomb matrix.

We studied nanoclusters of palladium, gold, platinum, and aluminum, binary clusters of Ru-Pt and Au-Pt, and ternary clusters of Ag-Au-Pt. The method is shown to produce all the reported global minimum structures, along with other minima, when we used the same or similar electronic structure methods and the same spin-states. Differences were seen when we used the semiempirical GFN-xTB method to compare the reported structure and properties from DFT or CCSD(T) studies. We have also evaluated some popular properties such as binding energy per atom, mixing energy, and compared with the reported ones.

We have varied some of the parameters in our approach for comparison of efficiency in finding the global minima and other properties of metal nanoclusters. We have compared different electronic structure methods, semiempirical and a few DFT functionals, in gold and aluminum clusters. While all the methods produced the same global minima for gold clusters, the geometries of maximum stability were highly dependent on the method for Al clusters. The method dependency was also seen in identifying the ground spin-state in Pt clusters. Thus, we can use less expensive methods such as semiempirical methods or empirical potentials for the clusters which do not change the ground-state multiplicity, and for which these methods give good results comparable to high-level *ab initio* or DFT methods. We can also use a two-layer approach where the initial search is done by cheaper methods, and the selected geometries can be optimized at a higher level.

We checked the effect of varying the number of orientations by comparing the binding energy per atom in the Au clusters. The study showed that as the cluster size increase more orientations has to be used for better results. In the same way, the result from different runs may vary if a small number of orientations are used, as was found by comparing the BE for three separate runs for Pt and Al clusters. As the cluster size increases, the search space increases and hence either number of orientations has to be increased or multiple runs have to be carried out to ensure that most of the local minima are found to increase the chance of finding the global minima.

A major potential challenge in such cluster-growing methods is the ability to capture the changes in structural motifs. We have seen that the GM's in Au_n clusters changed from 2D to 3D on going from n = 10–11. We also found similar structural changes in carbon clusters, 1D linear geometry, to 2D ring structures, and 3D structures including bowl-shaped geometries.

Thus, by building the cluster of size n by exploring three degrees of freedom involving the relative orientations of $n - 1$ and one atom gives the minima obtained by exploring 3N-6 degrees of freedom by the other methods. The direct comparison of complexity for the n-sized cluster is not meaningful because we have to add the complexity for exploring each of the clusters of size up to n. The major limitation our method is that it can be more expensive for building a cluster of a particular size of n as the method has to build all the clusters from size 2 to n. This method may not be very useful if one is only interested in only the n-sized cluster. Our method, however, is useful for studying the evolution of properties with growing cluster size.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the manuscript/**Supplementary Files**.

AUTHOR CONTRIBUTIONS

AA designed the project. MK and RM did the calculations. All the authors contributed to the manuscript preparation.

ACKNOWLEDGMENTS

We thank the Department of Science and Technology (DST), New Delhi, India for the FIST grants (SR/FST/CSII-026/2013 and SR/FST/CSII-011/2005) and IIT Kharagpur for computational facilities. MK thank DST for INSPIRE fellowship.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2019.00644/full#supplementary-material

REFERENCES

Adamo, C., and Barone, V. (1999). Toward reliable density functional methods without adjustable parameters: the PBE0 model. *J. Chem. Phys.* 110, 6158–6170. doi: 10.1063/1.478522

Ahlrichs, R., and Elliott, S. D. (1999). Clusters of aluminum, a density functional study. *Phys. Chem. Chem. Phys.* 1, 13–21. doi: 10.1039/a807713d

Anoop, A. (2019). *Pyar*. Available online at: https://github.com/anooplab/pyar

Arico, A., Srinivasan, S., and Antonucci, V. (2001). Dmfcs: from fundamental aspects to technology development. *Fuel Cells* 1, 133–161.

Assadollahzadeh, B., and Schwerdtfeger, P. (2009). A systematic search for minimum structures of small gold clusters aun (n=2-20) and their electronic properties. *J. Chem. Phys.* 131:064306. doi: 10.1063/1.3204488

Baek, H., Moon, J., and Kim, J. (2017). Benchmark study of density functional theory for neutral gold clusters, aun (n = 2-8). *J. Phys. Chem. A* 121, 2410–2419. doi: 10.1021/acs.jpca.6b11868

Balasubramanian, K. (1989). Ten low-lying electronic states of pd3. *J. Chem. Phys.* 91, 307–313. doi: 10.1063/1.457518

Baletto, F., and Ferrando, R. (2005). Structural properties of nanoclusters: energetic, thermodynamic, and kinetic effects. *Rev. Mod. Phys.* 77, 371–423. doi: 10.1103/RevModPhys.77.371

Barbati, M., Bruno, G., and Genovese, A. (2012). Applications of agent-based models for optimization problems: a literature review. *Exp. Syst. Appl.* 39, 6020–6028. doi: 10.1016/j.eswa.2011.12.015

Becke, A. D. (1988). Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* 38, 3098–3100. doi: 10.1103/PhysRevA.38.3098

Bion, N., Epron, F., Moreno, M., Marino, F., and Duprez, D. (2008). Preferential oxidation of carbon monoxide in the presence of hydrogen (prox) over noble metals and transition metal oxides: advantages and drawbacks. *Top. Catal.* 51:76. doi: 10.1007/s11244-008-9116-x

Bishea, G. A., and Morse, M. D. (1991). Spectroscopic studies of jet-cooled agau and au2. *J. Chem. Phys.* 95, 5646–5659. doi: 10.1063/1.461639

Dai, D., and Balasubramanian, K. (1995). Electronic structures of pd4 and pt4. *J. Chem. Phys.* 103, 648–655. doi: 10.1063/1.470098

Davis, J. B. A., Shayeghi, A., Horswell, S. L., and Johnston, R. L. (2015). The birmingham parallel genetic algorithm and its application to the direct DFT global optimisation of IrN(n = 10–20) clusters. *Nanoscale* 7, 14032–14038. doi: 10.1039/C5NR03774C

Demiroglu, I., Yao, K., Hussein, H. A., and Johnston, R. L. (2017). Dft global optimization of gas-phase subnanometer ru–pt clusters. *J. Phys. Chem. C* 121, 10773–10780. doi: 10.1021/acs.jpcc.6b11329

Doye, J. P. K., and Wales, D. J. (1998). Thermodynamics of global optimization. *Phys. Rev. Lett.* 80, 1357–1360. doi: 10.1103/PhysRevLett.80.1357

Eberhardt, W. (2002). Clusters as new materials. *Surf. Sci.* 500, 242–270. doi: 10.1016/S0039-6028(01)01564-3

Edwards, P. P., Johnston, R. L., Rao, C. N. R., Tunstall, D. P., and Johnston, R. L. (1998). The development of metallic behaviour in clusters. *Philos. Trans. R. Soc. Lond. Ser. A* 356, 211–230. doi: 10.1098/rsta.1998.0158

Fang, P.-P., Duan, S., Lin, X.-D., Anema, J. R., Li, J.-F., Buriez, O., et al. (2011). Tailoring au-core pd-shell pt-cluster nanoparticles for enhanced electrocatalytic activity. *Chem. Sci.* 2, 531–539. doi: 10.1039/C0SC00489H

Floudas, C. A., and Gounaris, C. E. (2008). A review of recent advances in global optimization. *J. Glob. Optim.* 45:3. doi: 10.1007/s10898-008-9332-8

Frisch, M., Trucks, G., Schlegel, H., Scuseria, G., Robb, M., Cheeseman, J., et al. (2016). Gaussian 16. *Revision A* 3.

Furche, F., Ahlrichs, R., Hättig, C., Klopper, W., Sierka, M., and Weigend, F. (2014). Turbomole. *Wiley Interdiscip. Rev.* 4, 91–100. doi: 10.1002/wcms.1162

Glover, F. (1986). Future paths for integer programming and links to artificial intelligence. *Comput. Operat. Res.* 13, 533–549. doi: 10.1016/0305-0548(86)90048-1

Glover, F. (1989). Tabu search—part i. *ORSA J. Comput.* 1, 190–206. doi: 10.1287/ijoc.1.3.190

Glover, F. (1990). Tabu search—part II. *ORSA J. Comput.* 2, 4–32. doi: 10.1287/ijoc.2.1.4

Götz, D. A., Schäfer, R., and Schwerdtfeger, P. (2013). The performance of density functional and wavefunction-based methods for 2d and 3d structures of au10. *J. Comput. Chem.* 34, 1975–1981. doi: 10.1002/jcc.23338

Götz, D. A., Shayeghi, A., Johnston, R. L., Schwerdtfeger, P., and Schäfer, R. (2016). Structural evolution and metallicity of lead clusters. *Nanoscale* 8, 11153–11160. doi: 10.1039/C6NR02080A

Grimme, S., Bannwarth, C., and Shushkov, P. (2017). A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all spd-block elements (z = 1-86). *J. Chem. Theory Comput.* 13, 1989–2009. doi: 10.1021/acs.jctc.7b00118

Grimme, S., Ehrlich, S., and Goerigk, L. (2011). Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* 32, 1456–1465. doi: 10.1002/jcc.21759

Gruene, P., Rayner, D. M., Redlich, B., van der Meer, A. F., Lyon, J. T., Meijer, G., et al. (2008). Structures of neutral au7, au19, and au20 clusters in the gas phase. *Science* 321, 674–676. doi: 10.1126/science.1161166

Heiles, S., and Johnston, R. L. (2013). Global optimization of clusters using electronic structure methods. *Int. J. Quant. Chem.* 113, 2091–2109. doi: 10.1002/qua.24462

Jäger, M., Schäfer, R., and Johnston, R. L. (2018). First principles global optimization of metal clusters and nanoalloys. *Adv. Phys. X* 3:1516514. doi: 10.1080/23746149.2018.1516514

James, A., Kowalczyk, P., Simard, B., Pinegar, J., and Morse, M. (1994). The a/1u ← x0+g system of gold dimer. *J. Mol. Spectrosc.* 168, 248–257. doi: 10.1006/jmsp.1994.1275

Jennings, P., and Johnston, R. (2013). Structures of small ti- and v-doped pt clusters: a GA-DFT study. *Comput. Theoret. Chem.* 1021, 91–100. doi: 10.1016/j.comptc.2013.06.033

Johnston, R. L. (2002). *Atomic and Molecular Clusters*. London: CRC Press.

Johnston, R. L. (2003). Evolving better nanoparticles: genetic algorithms for optimising cluster geometries. *Dalton Trans.* 2003, 4193–4207. doi: 10.1039/b305686d

Jones, R. (1993). Simulated annealing study of neutral and charged clusters: Al n and ga n. *J. Chem. Phys.* 99, 1194–1206. doi: 10.1063/1.465363

Jortner, J. (1992). Cluster size effects. *Zeitschrift für Physik D Atoms Molecules Clusters* 24, 247–275. doi: 10.1007/BF01425749

Joswig, J.-O., and Springborg, M. (2003). Genetic-algorithms search for global minima of aluminum clusters using a sutton-chen potential. *Phys. Rev. B* 68:085408. doi: 10.1103/PhysRevB.68.085408

Kaizuka, K., Miyamura, H., and Kobayashi, S. (2010). Remarkable effect of bimetallic nanocluster catalysts for aerobic oxidation of alcohols: combining metals changes the activities and the reaction pathways to aldehydes/carboxylic acids or esters. *J. Am. Chem. Soc.* 132, 15096–15098. doi: 10.1021/ja108256h

Khare, A., and Rangnekar, S. (2013). A review of particle swarm optimization and its applications in solar photovoltaic system. *Appl. Soft Comput.* 13, 2997–3006. doi: 10.1016/j.asoc.2012.11.033

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* 220, 671–680. doi: 10.1126/science.220.4598.671

Kittel, C. (2005). *Introduction to Solid State Physics*. New York, NY: John Wiley & Sons.

Lin, S.-S., Strauss, B., and Kant, A. (1969). Dissociation energy of pd2. *J. Chem. Phys.* 51, 2282–2283. doi: 10.1063/1.1672334

Liu, H., Song, C., Zhang, L., Zhang, J., Wang, H., and Wilkinson, D. P. (2006). A review of anode catalysis in the direct methanol fuel cell. *J. Power Sour.* 155, 95–110. doi: 10.1016/j.jpowsour.2006.01.030

Logsdail, A., Paz-Borbón, L. O., and Johnston, R. L. (2009). Structures and stabilities of platinum–gold nanoclusters. *J. Comput. Theor. Nanosci.* 6, 857–866. doi: 10.1166/jctn.2009.1118

López-Estrada, O. and Orgaz, E. (2015). Theoretical study of the spin competition in small-sized al clusters. *J. Phys. Chem. A* 119, 11941–11948. doi: 10.1021/acs.jpca.5b09871

Lv, J., Wang, Y., Zhu, L., and Ma, Y. (2012). Particle-swarm structure prediction on clusters. *J. Chem. Phys.* 137:084104. doi: 10.1063/1.4746757

Nandi, S., Bhattacharyya, D., and Anoop, A. (2018). Prebiotic chemistry of HCN tetramerization by automated reaction search. *Chemistry* 24, 4885–4894. doi: 10.1002/chem.201705492

Nandi, S., McAnanama-Brereton, S. R., Waller, M. P., and Anoop, A. (2017). A tabu-search based strategy for modeling molecular aggregates and binary reactions. *Comput. Theor. Chem.* 1111, 69–81. doi: 10.1016/j.comptc.2017.03.040

Nava, P., Sierka, M., and Ahlrichs, R. (2003). Density functional study of palladium clusters. *Phys. Chem. Chem. Phys.* 5, 3372–3381. doi: 10.1039/B303347C

Neese, F. (2018). Software update: the orca program system, version 4.0. *Wiley Interdiscip. Rev.* 8:e1327. doi: 10.1002/wcms.1327

Pacheco-Contreras, R., Juárez-Sánchez, J. O., Dessens-Félix, M., Aguilera-Granja, F., Fortunelli, A., and Posada-Amarillas, A. (2018). Empirical-potential global minima and dft local minima of trimetallic aglaumptn (l+ m+ n= 13, 19, 33, 38) clusters. *Comput. Mater. Sci.* 141, 30–40. doi: 10.1016/j.commatsci.2017.09.022

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. Available online at: http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

Perdew, J. P. (1986). Density-functional approximation for the correlation energy of the inhomogeneous electron gas. *Phys. Rev. B* 33, 8822–8824. doi: 10.1103/PhysRevB.33.8822

Pettersson, L. G., Wahlgren, U., and Gropen, O. (1983). Effective core potential calculations using frozen orbitals. applications to transition metals. *Chem. Phys.* 80, 7–16. doi: 10.1016/0301-0104(83)85164-7

Rao, B. K., and Jena, P. (1999). Evolution of the electronic structure and properties of neutral and charged aluminum clusters: a comprehensive analysis. *J. Chem. Phys.* 111, 1890–1904. doi: 10.1063/1.479458

Reveles, J. U., Köster, A. M., Calaminici, P., and Khanna, S. N. (2012). Structural changes of pd13 upon charging and oxidation/reduction. *J. Chem. Phys.* 136:114505. doi: 10.1063/1.3692612

Rogan, J., García, G., Valdivia, J. A., Orellana, W., Romero, A. H., Ramírez, R., et al. (2005). Small pd clusters: a comparison of phenomenological and ab initio approaches. *Phys. Rev. B* 72:115421. doi: 10.1103/PhysRevB.72.115421

Rupp, M., Tkatchenko, A., Müller, K.-R., and von Lilienfeld, O. A. (2012). Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* 108:058301. doi: 10.1103/PhysRevLett.108.058301

Sadeghi, A., Ghasemi, S. A., Schaefer, B., Mohr, S., Lill, M. A., and Goedecker, S. (2013). Metrics for measuring distances in configuration spaces. *J. Chem. Phys.* 139:184118. doi: 10.1063/1.4828704

Saha, K., Agasti, S. S., Kim, C., Li, X., and Rotello, V. M. (2012). Gold nanoparticles in chemical and biological sensing. *Chem. Rev.* 112, 2739–2779. doi: 10.1021/cr2001178

Schön, J., Putz, H., and Jansen, M. (1996). Studying the energy hypersurface of continuous systems-the threshold algorithm. *J. Phys.* 8:143. doi: 10.1088/0953-8984/8/2/004

Shayeghi, A., Götz, D., Davis, J. B., Schäfer, R., and Johnston, R. L. (2015a). Pool-BCGA: a parallelised generation-free genetic algorithm for the ab initio global optimisation of nanoalloy clusters. *Phys. Chem. Chem. Phys.* 17, 2104–2112. doi: 10.1039/C4CP04323E

Shayeghi, A., Götz, D., Johnston, R., and Schäfer, R. (2015b). Optical absorption spectra and structures of ag 6+ and ag 8+. *Eur. Phys. J. D* 69:152. doi: 10.1140/epjd/e2015-60188-2

Shayeghi, A., Heard, C. J., Johnston, R. L., and Schäfer, R. (2014). Optical and electronic properties of mixed ag-au tetramer cations. *J. Chem. Phys.* 140:054312. doi: 10.1063/1.4863443

Shi, L.-T., Wang, Z.-Q., Hu, C.-E., Cheng, Y., Zhu, J., and Ji, G.-F. (2019). Possible lower energy isomer of carbon clusters c (n = 11, 12) via particle swarm optimization algorithm: Ab initio investigation. *Chem. Phys. Lett.* 721, 74–85. doi: 10.1016/j.cplett.2019.02.028

Shi, Y.-K., Li, Z. H., and Fan, K.-N. (2010). Validation of density functional methods for the calculation of small gold clusters. *J. Phys. Chem. A* 114, 10297–10308. doi: 10.1021/jp105428b

Shinde, R., and Shukla, A. (2014). Large-scale first principles configuration interaction calculations of optical absorption in aluminum clusters. *Phys. Chem. Chem. Phys.* 16, 20714–20723. doi: 10.1039/C4CP02232G

Solov'yov, I. A., Solov'yov, A. V., and Greiner, W. (2004). Fusion process of lennard-jones clusters: global minima and magic numbers formation. *Int. J. Mod. Phys. E* 13, 697–736. doi: 10.1142/S0218301304002454

Song, C., Ge, Q., and Wang, L. (2005). Dft studies of pt/au bimetallic clusters and their interactions with the co molecule. *J. Phys. Chem. B* 109, 22341–22350. doi: 10.1021/jp0546709

Stewart, J. (2016). *Mopac 16*. Colorado Springs, CO: Stewart Computational Chemistry.

Turney, J. M., Simmonett, A. C., Parrish, R. M., Hohenstein, E. G., Evangelista, F. A., Fermann, J. T., et al. (2012). Psi4: an open-source ab initio electronic structure program. *Wiley Interdiscip. Rev.* 2, 556–565. doi: 10.1002/wcms.93

Wales, D. J., and Scheraga, H. A. (1999). Global optimization of clusters, crystals, and biomolecules. *Science* 285, 1368–1372. doi: 10.1126/science.285.5432.1368

Weigend, F., and Ahlrichs, R. (2005). Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for h to rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* 7, 3297–3305. doi: 10.1039/b508541a

Wen, J.-Q., Chen, G.-X., Zhang, J.-M., and Wu, H. (2018). The study of structures and properties of pdnhm(n = 1—10, m = 1,2) clusters by density functional theory. *J. Phys. Chem. Solids* 115, 84–91. doi: 10.1016/j.jpcs.2017.12.011

Xiao, C., Krüger, S., Belling, T., Mayer, M., and Rösch, N. (1999). Relativistic effects on geometry and electronic structure of small pdn species (n= 1, 2, 4). *Int. J. Quant. Chem.* 74, 405–416.

Zacarias, A. G., Castro, M., Tour, J. M., and Seminario, J. M. (1999). Lowest energy states of small pd clusters using density functional theory and standard ab initio methods. a route to understanding metallic nanoprobes. *J. Phys. Chem. A* 103, 7692–7700. doi: 10.1021/jp9913160

Zhang, J., and Dolg, M. (2015). Abcluster: the artificial bee colony algorithm for cluster global optimization. *Phys. Chem. Chem. Phys.* 17, 24173–24181. doi: 10.1039/C5CP04060D

Zhao, Y., Chen, X., and Li, J. (2017). Tgmin: a global-minimum structure search program based on a constrained basin-hopping algorithm. *Nano Res.* 10, 3407–3420. doi: 10.1007/s12274-017-1553-z

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 05 November 2019
doi: 10.3389/fchem.2019.00707

A New Genetic Algorithm Approach Applied to Atomic and Molecular Cluster Studies

Frederico T. Silva[1], *Mateus X. Silva*[2] *and Jadson C. Belchior*[3]*

[1] Departamento de Química Fundamental-CCEN, Universidade Federal de Pernambuco, Cidade Universitária, Recife, Brazil, [2] Programa de Pós-Graduação em Modelagem Matemática e Computacional, Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Belo Horizonte, Brazil, [3] Departamento de Química-ICEx, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

A new procedure is suggested to improve genetic algorithms for the prediction of structures of nanoparticles. The strategy focuses on managing the creation of new individuals by evaluating the efficiency of operators (o_1, o_2,...,o_{13}) in generating well-adapted offspring. This is done by increasing the creation rate of operators with better performance and decreasing that rate for the ones which poorly fulfill the task of creating favorable new generation. Additionally, several strategies (thirteen at this level of approach) from different optimization techniques were implemented on the actual genetic algorithm. Trials were performed on the general case studies of 26 and 55-atom clusters with binding energy governed by a Lennard-Jones empirical potential with all individuals being created by each of the particular thirteen operators tested. A 18-atom carbon cluster and some polynitrogen systems were also studied within REBO potential and quantum approaches, respectively. Results show that our management strategy could avoid bad operators, keeping the overall method performance with great confidence. Moreover, amongst the operators taken from the literature and tested herein, the genetic algorithm was faster when the generation of new individuals was carried out by the twist operator, even when compared to commonly used operators such as Deaven and Ho cut-and-splice crossover. Operators typically designed for basin-hopping methodology also performed well on the proposed genetic algorithm scheme.

Keywords: cluster optimization, quantum genetic algorithm (QGA), evolutionary operator management, Lennard-Jones clusters, polynitrogen structure optimization

OPEN ACCESS

Edited by:
Jorge M. C. Marques,
University of Coimbra, Portugal

Reviewed by:
Francisco Pereira,
Instituto Politécnico de Coimbra, Portugal
William Tiznado,
Universidad Andrés Bello, Chile

****Correspondence:***
Jadson C. Belchior
jadson@ufmg.br

Specialty section:
This article was submitted to Theoretical and Computational Chemistry,
a section of the journal Frontiers in Chemistry

Received: *05 June 2019*
Accepted: *09 October 2019*
Published: *05 November 2019*

Citation:
Silva FT, Silva MX and Belchior JC (2019) A New Genetic Algorithm Approach Applied to Atomic and Molecular Cluster Studies. Front. Chem. 7:707. doi: 10.3389/fchem.2019.00707

1. INTRODUCTION

Clusters are aggregates of atoms or molecules whose structures remain between those of discrete atoms and of the bulk material (Johnston, 2003). Moreover, their properties are composition and size dependent. Palladium, for instance, is non-magnetic in the solid state, but its counterpart clusters may have non-zero magnetic moment (Moseler et al., 2001). Among the wide range of interesting cluster applications one could mention magnetic resonance imaging (Lu et al., 2017), water oxidation (Zhao et al., 2017), magnetic storage (Bader, 2006), and catalysis (Pelegrini et al., 2016). In addition, clusters are promising in the development of nanomachines (Rieth and Schommers, 2002), Islas et al. and Merino et al., for example, showed the stability of boron wheels (Islas et al., 2007; Jiménez-Halla et al., 2010), while the latter researchers also studied the aromaticity of such particles and the rotational motion of these rings with respect to each other, comparing their behavior to a wankel motor (Jiménez-Halla et al., 2010). However, for most of computational chemistry techniques, atomic coordinates are needed for the calculation of clusters properties, and hence one must know the cluster structure. Finding the geometries of small clusters is a challenging task and requires a combination of theoretical and experimental techniques (Götz et al., 2012; Heiles et al., 2012).

It is generally assumed that clusters adopt the lowest energy structure (Lazauskas et al., 2017). Accordingly, finding such structure is a matter of finding the global minimum of an appropriate potential energy surface (PES). Modeling such PES within a quantum approach rapidly becomes computationally prohibitive, therefore empirical analytic expressions are usually employed to describe the interactions between the particles composing the clusters. Examples of these potentials are the Lennard-Jones (Jones and Ingham, 1925), Morse (Morse, 1929), and REBO (Brenner et al., 2002; Kosimov et al., 2010; Bonnin et al., 2019; Jiang et al., 2019; Lin et al., 2019) potentials, the latter being a more complex one which has gained prominence due to its applicability to describe graphene potential energy surfaces (Jiang et al., 2019; Lin et al., 2019).

Once the way to compute the energy of the system has been defined, one must minimize it. There are several techniques that enable global minima search, such as big bang methodology (BB) (Lazauskas et al., 2017), basin-hopping (BH) (Rondina and Da Silva, 2013) and evolutionary algorithms, such as genetic algorithms (GA) (Johnston, 2003). Especially, GAs have been successfully applied to predict chemical structures from clusters to protein folding (Johnston, 2003; Louis and McDonnel, 2004; Heiles et al., 2012; Silva et al., 2014b; Borguesan et al., 2015; Song et al., 2018). Even so, finding the global minimum associated with these chemical systems implies efficiently exploring the most reasonable portions of their PES, which still is a challenging task. Therefore, new algorithms are constantly being developed.

2. RELATED WORK

It is already well discussed in the literature that, in order to guarantee efficiency in convergence and appropriate exploration of the PES associated with atomic and molecular clusters, evolutionary algorithms employed in global optimization problems must ensure population diversity (Hartke, 1999; Cheng et al., 2004; Grosso et al., 2007; Pereira and Marques, 2009; Marques et al., 2018). Therefore, estimating how similar are the structures composing the evolving population can provide valuable information to assist the evolutionary procedure. In the work of Hartke (1999), it is proposed that a minimum degree of exploration of the PES is ensured by making part of the population always composed by mutants. That means a set of structures that have been randomly modified will be present throughout the evolutionary procedure, regardless of whether they are better adapted or not. In the same work, a minimum energy difference between structures is established to maintain diversity, as well as a balance between optimization performance and exploration of the PES is proposed through the simultaneous use of a random operator such as Deaven and Ho (1995) cutting plane and a biased version of this operator in which the cluster is separated into its best and worst halves. Hartke (1999) also proposes a measure based on the two-dimensional projections of clusters structures that can distribute different types of geometries into niches. Thus, different ranges of values can be assigned to different types of geometries, allowing the evaluation of structure similarities and enabling one to avoid population stagnation.

Cheng et al. (2004) propose that structure similarity checking should always be based on topological information, and that measurements of the distance between energy minimum structures should be carried out by comparing numerical values associated with structure similarities. In their work, a connectivity table for cluster similarity checking is proposed, in which the connectivity information of a cluster is characterized according to the number of atoms having *i* nearest neighbors within the cluster. By using this connectivity table together with the evaluation of the fitness of each individual, they managed to balance diversity and convergence efficiency. Pereira and Marques (2009) state that one should consider structural information for estimating dissimilarities among cluster structures when searching for energy minima within an evolutionary algorithm approach, instead of taking into account fitness values. They have employed a combination of an evolutionary approach with a local search method that uses derivative information to search for the nearest local minimum without requiring any previous knowledge about the problem being solved. The authors show that maintaining diversity is the main issue to guarantee effectiveness, which was carried out by the application of three distinct distance measures to estimate the dissimilarity between structures.

As for recent advances in the development of genetic algorithms, Heiles et al. coupled Plane-Wave Self-Consistent Field (PWscf) package with Birmingham Cluster Genetic Algorithm (BCGA), allowing the study of Au-Ag nanoalloys through density functional theory (Heiles et al., 2012). Zayed et al. implemented what they called universal genetic algorithm, making use of Python's large collection of libraries and of the scaling capabilities of a pool genetic algorithm (Zayed et al., 2017). Vilhelmsen and Hammer proposed an inexpensive strategy to eliminate similar structures from the population (Vilhelmsen and Hammer, 2012). Lazauskas et al. proposed a pre-screening to eliminate structures with high probability of convergence failure during local minimization (Lazauskas et al., 2017).

In the past we proposed two new operators, namely annihilator and history operators (Guimarães et al., 2002), that demonstrated along the years (Lordeiro et al., 2003; Rodrigues et al., 2008; Silva et al., 2014a,b) to be quite efficient for determining global minima in atomic and molecular cluster studies where many local minima were present. Regarding the creation of new individuals, one can observe a broad variation among methodologies available in the literature. In general, each operator application rate is kept constant throughout the GA execution. For instance, Wang et al. used the values 0.5, 0.3, and 0.2 for mating, mutation and exchange rates, respectively, in their global minimization (Wang et al., 2018). Zhao et al. propose values between 10% and 30% for mutation rate (Zhao et al., 2016), while in an outline of the evolutionary principles of GAs, Heiles and Johnston describe a parameter that defines the probability of mutation, p_{mut} (Heiles and Johnston, 2013). Let n_{tot} be the total number of individuals to be created after energy minimization of an arbitrary generation; among them, $p_{mut}n_{tot}$ individuals are created by mutation operators, while $(1-p_{mut})n_{tot}$ are created by crossover or recombination methods, on average (Heiles and Johnston, 2013). Finally, Rondina *et al.* used a dynamic strategy to manage operators in a basin-hopping technique (Rondina and Da Silva, 2013).

In this work, we propose a method with dynamic management of evolutionary operators for genetic algorithms that, in principle, could lead to a more efficient way to survey the PES of atomic and molecular clusters than our previous older GA version (Guimarães et al., 2002; Lordeiro et al., 2003). The paper is divided as follows: section 3 outlines a standard GA procedure, gives the details of our algorithm and describes all the operators employed as well as the management strategy proposed. The comparison between the different builds tested and the evaluation of their behavior according to the model system employed are presented and discussed in section 4. The main conclusions are gathered in section 5.

3. METHODOLOGY

3.1. Genetic Algorithm Procedure

A standard GA procedure is defined by three main steps. The first step is initialization, when an initial population of individuals is generated. The second step is selection, where all individuals are ranked according to their fitness and, in the present work, the 25% worst are eliminated. The third step is the creation of new individuals, where, in general, operators are applied to individuals that survived the selection step to generate new structures. We call operators all ways of generating a new member of the population. Desirable operators are the ones which efficiently span the potential energy surface of the system representatively. This can be done in different ways to which different concepts are associated and will be discussed further on. After creation step, the whole population is submitted to selection again and the cycle is repeated (Johnston, 2003).

One can find a wide variety of genetic algorithms in which the basic structure just described has been customized to improve performance or to meet some specific needs. In fact, the generation of the initial population may not always be completely random (Johnston, 2003; Chen et al., 2013); the measure of the quality of the individuals (fitness) might be given by different mathematical approaches (Burton and Vladimirova, 1998; Jin et al., 2002; Yan and Wang, 2010), and its upper and lower limits may be fixed or scaled in each generation according to the current population (Johnston, 2003). The selection of individuals to be eliminated or to generate offspring may depend on their fitness values in different ways, as well as various methods are available for choosing parents for mating (Saini, 2017). Furthermore, subpopulations can be evolved in parallel and exchange individuals along the procedure, simulating migration in natural populations (Chen et al., 2013). These few examples, and all their possible combinations, illustrate the versatility of genetic algorithms.

In this work, however, we concentrate mainly on the creation of new individuals within an approach focused on the study of atomic and nanoalloy clusters. Our approach changes the creation rate of each operator employed on the fly, favoring the better ones. In order to do so, we first performed a study over 13 evolutionary operators collected in the literature (Deaven and Ho, 1995; Michalewicz, 1996; Wales and Doye, 1997; Johnston, 2003; Takeuchi, 2007; Kim et al., 2008; Ye et al., 2011; Chen et al., 2013; Rondina and Da Silva, 2013) and evaluated the performance of all proposed builds within a 26 and 55-atom Lennard-Jones potential clusters (LJ_{26} and LJ_{55}) approach and a simple evolutionary scheme.

We have employed a primary GA framework and focused on the outcomes of each operator, both individually tested and jointly implemented, when tackling simple systems such as LJ_{26} and LJ_{55}. We have also briefly approached the harder LJ_{38} system, the C_{18} cluster employing the more complex REBO potential and applied our management strategy within a quantum approach to polynitrogen systems. The scheme of the genetic algorithm implemented in this work is presented in **Figure 1**, and each of its steps will be discussed in the following sessions. The program was written in C++ and the calculations were made on an Unix computer. For the polynitrogen cases, however, a more robust algorithm (Silva et al., 2018) was chosen (coupled to GAMESS-US, Schmidt et al., 1993). In the future we intend to both extend this approach to molecular nanoclusters and enhance the efficiency of our algorithm by improving each of its steps with typical strategies (Johnston, 2003) that help avoiding unnecessary computational effort and assist convergence.

3.2. Initialization

Following Cai et al. (2002), each individual of the initial population is created by randomly generating atoms inside a mathematical sphere of radius R, defined by Equation (1):

$$R = r_e\left[\frac{1}{2} + \left(\frac{3N}{4\pi\sqrt{2}}\right)^{\frac{1}{3}}\right] \tag{1}$$

where N is the number of atoms and r_e is a parameter related to the equilibrium distance between atoms, here set to 1.0. Additionally, a restriction was added to prevent atoms from being generated very close to each other. The minimum distance allowed between two atoms at this step is 0.8 (dimensionless units adopted).

3.3. Selection and Stop Condition

In the present work, two analytic potentials were chosen to define the potential energy surfaces to be explored, namely the Lennard-Jones and REBO potentials. The adjustment of the parameters

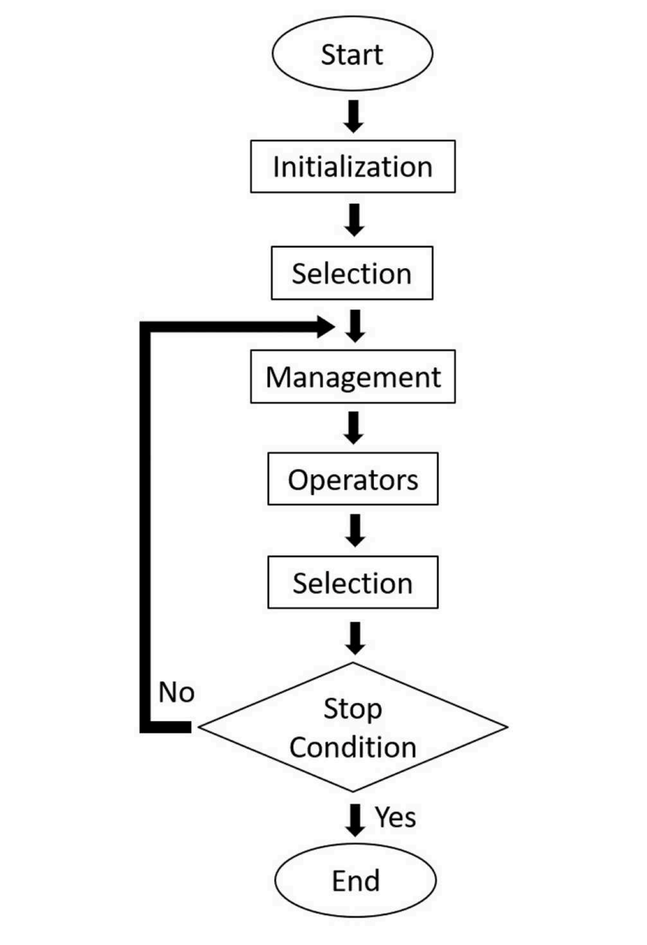


FIGURE 1 | A flowchart of our genetic algorithm procedure.

associated with the operators tested, as well as the evaluation of the employed builds were carried out using Lennard-Jones empirical potential (Equation 2) with reduced units ($\epsilon = \sigma = 1$).

$$E = 4\epsilon \sum_{i<j} \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right] \tag{2}$$

REBO potential was used further on to test the ability of the proposed builds to reach the global minimum in a more complex problem, the C_{18} cluster. This potential is described in detail in Brenner et al. (2002) and Kosimov et al. (2010), and it has been implemented with support from the Atomic Simulation Environment (ASE) library (Larsen et al., 2017). Among the options available in this simulation environment, we opted for the implementation of REBO present in Atomistica library. We have used dlib (King, 2009) library with limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm for local minimizations. In each generation, individuals were sorted by potential energy and the 25% worst were eliminated. From the remaining population, parent individuals are chosen for mating employing an uniform selection method, in which they are selected randomly uniformly from the current population. Individuals are also selected for mutation in the same manner. Although each operator has its own probability of acting over the population in each generation, it governs only the operator creation rate, but not the choice of an specifically ranked individual to act on.

Two stopping criteria were defined for the developed genetic algorithm. The first one is satisfied when the global minimum is achieved, which is already known for the Lennard-Jones 26, 38, and 55-atom cases studied here (−108.315616, −173.928427, and −279.248470, respectively), whose structures are shown in **Figure 2**. Also, the most stable structure for the 18-atom carbon cluster is known to be a planar monoring (shown in **Figure 3**), with binding energy equal to −108.3726 eV (Kosimov et al., 2010). The other criterion is fulfilled when 3,000 local minimizations are performed for C_{18} and the smaller Lennard-Jones cluster, and 5,000 local minimizations for the larger Lennard-Jones clusters. Usually, the global minimum is not known, and thus another termination criterion must be defined. For the polynitrogen cases tested, for example, the procedure is stopped either if it reaches 400 generations or if an individual remains as the one with the lowest energy for 20 consecutive generations (Silva et al., 2018). However, the former described stopping criteria are suitable for this work because performance was evaluated according to the number of local minimizations (N_{LM}) needed to reach the global energy minimum. Therefore, after reaching this point (which is already known), additional calculations are not necessary. This performance assessment suggested here was also used by Chen et al. (2013) for the proposition of a new crossover operator, where a sphere is used to cut and splice the parent structures, rather than a plane.

3.4. Management

The method we propose to manage the application of operators within the evolutionary procedure is based on setting, on the fly, the creation rate of each operator employed according to their outcomes. When a new individual is generated, its energy is compared to the average energy of the entire population that survived the previous selection step. If the energy of the new individual is lower than this average energy, more individuals will be created with that operator in the next generation. If, on the other hand, the energy of the new individual is above that average, the related operator suffers a decrease in its creation rate. The function chosen to describe how the creation rate of each operator o_j changes along the evolutionary procedure (from the current to the next generation) is piecewise-defined:

$$\upsilon_{ij}\left(\Delta E_{ij}\right) = \begin{cases} \upsilon_{max}, & \Delta E_{ij} < -\Delta E_{max} \\ \left(\frac{\upsilon_{max}}{\Delta E_{max}}\right) \Delta E_{ij}, & -\Delta E_{max} < \Delta E_{ij} < \Delta E_{max} \\ -\upsilon_{max}, & \Delta E_{ij} > \Delta E_{max} \end{cases} \tag{3}$$

where υ_{ij} is the i^{th} contribution to the variation of the creation rate of operator j. ΔE_{ij} is the energy difference between the new individual i, created by o_j, and the average of the population that survived the previous selection step. The maximum allowed value for the variation of the creation rate of any operator employed in

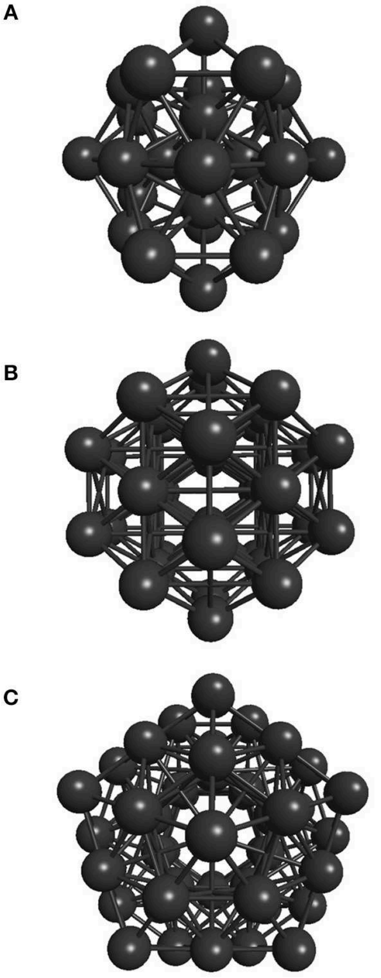


FIGURE 2 | Well-known global energy minimum structures of the **(A)** 26-atom, **(B)** 38-atom, and **(C)** 55-atom Lennard-Jones clusters.

the algorithm was defined as υ_{max} and here set to 0.9. ΔE_{max}, here set to 2.0, is the energy difference that triggers maximum variation. Thus, the new creation rate (p_j) of each operator o_j is defined according to Equation (4):

$$p_j = p'_j + \frac{1}{N}\sum_{i}^{N} \upsilon_{ij}(\Delta E_{ij}) \tag{4}$$

where p'_j is the creation rate of operator j in the previous generation, and N is the number of individuals it has created in the current generation.

After all creation rates have been modified, their sum is normalized to one. Lastly, new individuals are created by the rule $p_j n_{tot}$, where p_j is the creation rate of operator j and n_{tot} is total number of individuals that must be created.

3.5. Operators

Traditionally, the creation of new individuals is done in two different manners: through crossover or mutation (Johnston, 2003). Crossover combines two individuals from the population to produce new ones, simulating the combination of genetic information from the parents to generate offspring. Mutation modifies the coordinates of a single individual from the population to generate a new one, avoiding population stagnation. It simulates the introduction of new genetic material to the population. In this work, three types of operators are used: crossover, which produces a new individual combining other two; mutation, which produces a new individual from a single one; and immigration, which creates a new individual from scratch, simulating migration in natural environment.

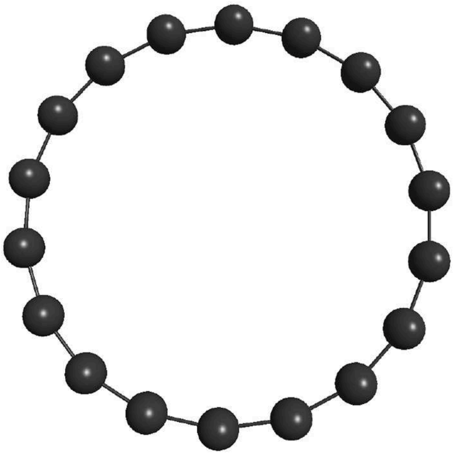


FIGURE 3 | Most stable structure for C_{18} cluster: a planar single-ring (Kosimov et al., 2010).

The following operators are of crossover type. They take two individuals (k and l) from the remaining population, chosen randomly from the group of the previous selection step survivors, to generate a new one (m).

a. Arithmetical crossover (ARCR) (Michalewicz, 1996): let $\boldsymbol{x}^{(k)}$ be the cartesian coordinate vector of individual k, $\boldsymbol{x}^{(l)}$ the cartesian coordinate vector of individual l and $\boldsymbol{x}^{(m)}$ the cartesian coordinate vector of the new cluster m. ARCR acts to generate a new individual with the following rule: $\boldsymbol{x}^{(m)} = 0.5(\boldsymbol{x}^{(k)} + \boldsymbol{x}^{(l)})$.
b. Plane-cut-splice crossover (PCCR) (Deaven and Ho, 1995): a plane is randomly defined separating the atoms of cluster k into two groups. Another random plane is defined for cluster l, also separating its atoms into two groups. The groups generated from cluster k must have the same number of atoms of those generated from cluster l. Then, equivalent groups are exchanged between clusters k and l to generate the new cluster m with the correct number of atoms.
c. Sphere-cut-splice crossover (SCCR) (Chen et al., 2013): analogous to PCCR, but using a sphere instead of a plane. A mathematical sphere is defined to separate cluster k into two groups of atoms, one that lies in the inner part of the sphere and other that lies in its outer region. The same sphere is generated for cluster l. If the inner part of both

progenitor clusters contains the same number of atoms, they are interchanged to generate a new individual m.

d. Two points crossover (TWCR) (Johnston, 2003): the coordinates of the atoms composing each of the selected individuals for mating must be arranged in a one-dimensional array. Then, two random integers are generated: $s_1 = [1, (3N - 1)]$ and $s_2 = [(s_1 + 1), 3N]$. The notation $[a, b]$ means that a random number between a and b, in a uniform distribution, must be generated. The coordinates of cluster k that lie between s_1 and s_2 are replaced by those of cluster l that lie on the same range.

e. Uniform crossover (UNCR) (Johnston, 2003): when generating the new individual $\boldsymbol{x^{(m)}}$, each new coordinate ($\boldsymbol{x_i^{(m)}}$) has a specific probability of coming from each of its parents. In the present approach the new individual has 70% chance of coming from cluster k ($\boldsymbol{x_i^{(k)}}$) and 30% chance of coming from cluster l ($\boldsymbol{x_i^{(l)}}$).

The following operators are of mutation type. They take one individual, k, also chosen randomly from selection step survivors, to generate a new one, m.

a. Angular operator (AO) (Wales and Doye, 1997): this operator acts on 1–5% of the total number of atoms in the cluster, chosen randomly. Each selected atom is displaced randomly over the surface of a sphere of radius R_i (equal to the distance of the atom to the geometric center of the cluster) centered in the geometric center of the particle.

b. Cartesian displacement operator (CDO) (Rondina and Da Silva, 2013): this operator acts on 1 to N atoms, chosen randomly. N is the total number of atoms in the cluster. Each selected atom is modified by the following equation:

$$\boldsymbol{r_i^{(m)}} = \boldsymbol{r_i^{(k)}} + S\, r_{min}(\,[\,-1, +1\,]\,\hat{\mathbf{\imath}} + [\,-1, +1\,]\,\hat{\mathbf{\jmath}} + [\,-1, +1\,]\,\hat{\mathbf{k}}) \tag{5}$$

where $\boldsymbol{r_i^{(m)}}$ are the new coordinates of the cluster's i^{th} atom, $\boldsymbol{r_i^{(k)}}$ are the former coordinates of that same atom, S is an arbitrary parameter, here set to 0.2, $\hat{\mathbf{\imath}}$, $\hat{\mathbf{\jmath}}$ and $\hat{\mathbf{k}}$ are the cartesian unit vectors, and r_{min} is the distance to the nearest atom to which the operator will act. Again, the notation $[\,-1, +1\,]$ means that a random number with uniform distribution must be generated between -1 and $+1$.

c. Dynamic mutation (DYM) (Johnston, 2003): this operator acts on all atoms of the selected individual according to the following equation:

$$\boldsymbol{r_i^{(m)}} = [\,(\,1 - \delta\,), (\,1 + \delta\,)\,]\,\boldsymbol{r_i^{(k)}} \tag{6}$$

where $\boldsymbol{r_i^{(m)}}$ are the new coordinates of the cluster's i^{th} atom, $\boldsymbol{r_i^{(k)}}$ are the former coordinates of that same atom and δ is an arbitrary parameter, here set to 0.10.

d. Geometric center displacement operator (GCDO) (Kim et al., 2008; Rondina and Da Silva, 2013): this operator acts on 1 to N atoms, chosen randomly. N is the total number of atoms in the cluster. Each selected atom is modified by the following equation:

$$\boldsymbol{r_i^{(m)}} = \boldsymbol{r_i^{(k)}} + [\,(\,\alpha_{max} - \alpha_{min}\,)\left(\frac{R_i}{R_{max}}\right)^{w} + \alpha_{min}\,]\,r_{min}\,\hat{e}_i\,(\,\theta_i, \varphi_i\,) \tag{7}$$

where $\boldsymbol{r_i^{(m)}}$ are the new coordinates of the cluster's i^{th} atom, $\boldsymbol{r_i^{(k)}}$ are the former coordinates of that same atom, r_{min} is the distance between the i^{th} atom and its nearest neighbor, R_i is the distance between the i^{th} atom and the geometric center of the particle, R_{max} is the distance between the center of the particle and its furthest atom, α_{max}, α_{min} and w are arbitrary parameters, here set to 0.2, 0.7, and 2.0, respectively, and $\hat{e}_i(\theta_i, \varphi_i)$ is a unit vector generated randomly in a spherical distribution.

e. Interior operator (IO) (Takeuchi, 2007; Ye et al., 2011): this operator moves a single atom toward the particle's nucleus. Let R_i be the distance between the i^{th} atom and the geometric center of the particle. Atom i is moved to a random position on the surface of a sphere of radius $[\,0.01, 0.10\,]R_i$, centered on the geometric center of the particle.

f. Surface angular operator (SAO) (Ye et al., 2011): this operator moves a single atom toward the surface of the cluster. Let R_{max} be the distance between the geometric center of the particle and its furthest atom. Selected atom, i, is moved to a random position on the surface of a sphere with radius R_{max}, centered on the geometric center of the particle.

g. Twist operator (TO) (Johnston, 2003; Rondina and Da Silva, 2013): a random plane is defined to separate the selected cluster into two portions, not necessarily with the same sizes. Then, one of these portions is rotated randomly around the axis formed by the normal to that plane. In this work, the angle of rotation, θ, was generated randomly between 0.10π and 0.50π.

The following operators are of immigration type. They create a new individual, m, from scratch.

a. Immigration (IMM and IMM0) (Cai et al., 2002): this operator generates a new individual in the same manner the initial population is created. Namely, atoms are generated randomly inside a sphere of radius defined by Equation (1). Two types of immigration are defined: IMM and IMM0. IMM has a restriction that prevents atoms from being created closer than 0.8 Å to each other. IMM0 does not have any restriction.

3.6. Test Methodology

In order to implement this new methodology for a GA based on the management of various mathematical operators, several builds were designed using the operators just described, individually and combined. Several tests were performed as well. All tests followed the same protocol in which the genetic algorithm was executed 50 times with different random number seeds. As described in section 3.3, the chosen systems were the general case studies of 26 and 55-atom clusters with binding energy governed by a Lennard-Jones empirical potential (LJ$_{26}$

and LJ_{55}) with $\epsilon = \sigma = 1$ (reduced units). Two stopping conditions were used: finding the global minimum or reaching 3,000 local minimizations for LJ_{26} or 5,000 local minimizations for LJ_{55}. The cluster population was kept constant in 40 individuals, 10 of them being eliminated at each generation and replaced by the available creation operators. The initial population was randomly generated using the method described in section 3.2.

Three different builds were proposed and used to test the management methodology described in section 3.4, namely AUTO5, AUTO7, and AUTO13. The numbers indicate how many creation operators were employed in each build. AUTO5 is composed by the following operators: TO, IO, PCCR, SAO, and IMM. AUTO7 is composed by: TO, IO, PCCR, SAO, IMM, AO, and GCDO. Finally, AUTO13 build is composed by all the operators tested herein: TO, IO, PCCR, SAO, IMM, AO, GCDO, TWCR, CDO, SCCR, UNCR, ARCR, and DYM. We have also run the same build of our previous work, here named PREV, which is composed of 70% SCCR, 20% DYM, and 10% IMM, kept fixed throughout the GA execution (Silva et al., 2015). Operator acronyms were defined in section 3.5.

TABLE 1 | Results of tests performed on our GA builds for LJ_{26}.

Build	$\widehat{N_{LM}}$*	$\sigma_{\bar{x}}$	N_{fails}(%)
TO	186	17	0
IO	205	16	0
PCCR	246	32	0
AUTO5	246	39	0
AUTO7	264	27	0
SAO	264	23	0
IMM	297	39	0
AO	272	33	2
AUTO13	434	61	2
GCDO	357	66	4
PREV**	720	120	6
UNCR	1,096	135	12
TWCR	634	62	16
CDO	739	113	16
SCCR	371	71	28
IMM0	1,242	197	52
ARCR	390	71	80
DYM	3,001	0	100

$\widehat{N_{LM}}$ indicates the average number of local minimizations needed to reach the global minimum. $\sigma_{\bar{x}}$ indicates the standard error and N_{fails} indicates the relative number of times the global minimum was not reached.

**The unconverged runs were removed from the calculation of these averages. This removal may compromise the analysis when N_{fails} is nonzero.*

***Previous work Silva et al. (2015).*

4. RESULTS AND DISCUSSION

Results yielded by all builds tested are presented in **Table 1** for the LJ_{26} case. $\widehat{N_{LM}}$ is the average number of local energy minimizations needed to achieve convergence to the global minimum, $\sigma_{\bar{x}}$ is the standard error, defined as the standard deviation divided by the square root of the total number of samples, and N_{fails} is the percentage of seeds employed that did not achieve convergence for a specific build. For those cases that failed to converge N_{LM} was defined as the maximum allowed number of local minimizations plus one (3001). However, unconverged runs were not taken into account to obtain $\widehat{N_{LM}}$. The results are presented primarily in ascending order of N_{fails}; secondarily, in ascending order of $\widehat{N_{LM}}$.

We call attention to builds TO, IO, PCCR, AUTO5, AUTO7, SAO, and IMM, which managed to find the global minimum on every run. On the other hand, DYM was the only one that failed to properly converge on all test runs with different values assigned to the δ parameter. Acting on all atoms of the selected individual at once seems to be an ineffective mutation for our purpose.

Twist operator (TO) was the one with lowest $\widehat{N_{LM}}$, however, it overlaps with interior operator (IO) if we take their standard errors into account. Within the same analysis, standard errors show that IO performed similarly to PCCR and AUTO5, which in turn were essentially equivalent to AUTO7 and SAO. Since the global minimum of LJ_{26} is approximately of spherical shape, it favored interior (IO) and surface angular (SAO) operators, explaining their good performances. In order to compare our top ranked build (TO) with the widely used plane-cut-splice (PCCR), which did not overlap considering their standard errors, we have used one-tailed p-value approach (Chaubey, 1993) and calculated that the twist operator build was better than plane-cut-splice crossover build with a 90% confidence level. To ensure that this comparison would be valid, we had previously tested for the normality of the data generated by these builds using the Ryan-Joiner test (Yap and Sim, 2011), and the normality hypothesis was accepted within a significance level of 0.01 with less than five percent of discrepant data removed. The PCCR proposed by Deaven and Ho (1995), however, still had a good performance, since its build managed to find the global minimum on every run and presented one of the lowest $\widehat{N_{LM}}$ values. This operator is employed in most of modern genetic algorithms (Johnston, 2003; Heiles and Johnston, 2013) and had its robustness already reevaluated, showing good results (Froltsov and Reuter, 2009).

The geometric center displacement operator (GCDO) presented better performance than the cartesian displacement operator (CDO). The parameters associated with each of these methods were refined before final test in both cases. The better GCDO performance could be explained by the two additional parameters available for tuning compared to CDO. The uniform crossover (UNCR), two points crossover (TWCR) and arithmetical crossover (ARCR) were not originally developed for cluster studies, and, among them, TWCR was the one that presented the best performance. They make up the worst performing group within the LJ_{26} approach along with CDO, SCCR, IMM0, and DYM.

Still for the LJ_{26} case, the sphere-cut-splice crossover (SCCR) performed poorly, which was expected since Chen et al. indeed reported that this operator is more suitable for larger clusters (Chen et al., 2013). In our previous work (Silva et al., 2015), the employed build (PREV) was mainly composed by SCCR, but also counted with the immigration operator and a different evolutionary scheme. Within the present GA approach, our PREV build presented worse performance ($\widehat{N_{LM}} = 720$) than SCCR-only build ($\widehat{N_{LM}} = 371$) when it comes to the average number of local minimizations needed to reach global minimum. Nevertheless, the PREV build presented better reliability than SCCR-only on finding the correct cluster structure, since the failure rate of the latter (28%) was almost five times greater than that of the former (6%). If we took the unconverged runs into consideration to calculate $\widehat{N_{LM}}^*$, PREV would go from $\widehat{N_{LM}} = 720$ to $\widehat{N_{LM}}^* = 857$, while SCCR would go from $\widehat{N_{LM}} = 371$ to $\widehat{N_{LM}}^* = 1107$. The improvement of our PREV build over the SCCR-only could be explained by the joint presence of the IMM operator in the PREV build, which had better performance here and was responsible for the creation of 10% of n_{tot} in each generation in our previous work. The immigration operator (IMM) itself was the fifth most efficient in the present study. However, it is important to note that the restriction that prevents atoms from being created too close to each other was decisive in its performance. Without such restriction, $\widehat{N_{LM}}$ goes from 297 (IMM) to 1242 (IMM0). Besides, IMM0 failed to converge on 52% of the trials within the LJ_{26} approach.

In our PREV build we were focused on the development of a GA to be coupled with electronic structure methods. Therefore, we needed to generate reasonable structures from the very beginning, while keeping population diversity within an unbiased analysis. That is because bad structures may easily lead to unconverged energy calculations or local minimizations in a quantum approach, unlike the empirical potential case, in which the energy may always be obtained analytically. On the other hand, avoiding completely stochastic contributions to the evolutionary procedure could prevent us from finding new energy minima, typically hard to guess if one has no previous information about the system. Seeking for good cost-effectiveness relation was essential to survey the *ab initio* potential energy surface associated with atomic clusters without calling upon empirical potentials. However, that was a difficult task to fulfill employing specific operators with fixed creation rates.

Exploring different possibilities of combining these evolutionary operators together may provide more flexibility to the algorithm and hence allow more thorough sampling of the PES in a single run. In the first instance, we are mostly interested in evaluating solely the contribution of the operators to the general performance of genetic algorithms. From this perspective, we can evaluate the behavior of our AUTO5, AUTO7, and AUTO13 builds within the highly unbiased GA scheme adopted here, in which the simplest rules were used to generate the population, to rank it and to select individuals for mating and mutation, as well as for predation. Through the graphs presented in **Figures 4–6**, we can assess the variations in the creation rate of each operator within our management strategy along different GA runs (chosen randomly) concerning the LJ_{26} system. **Figure 4** refers to AUTO5, **Figure 5** to AUTO7 and **Figure 6** to AUTO13.

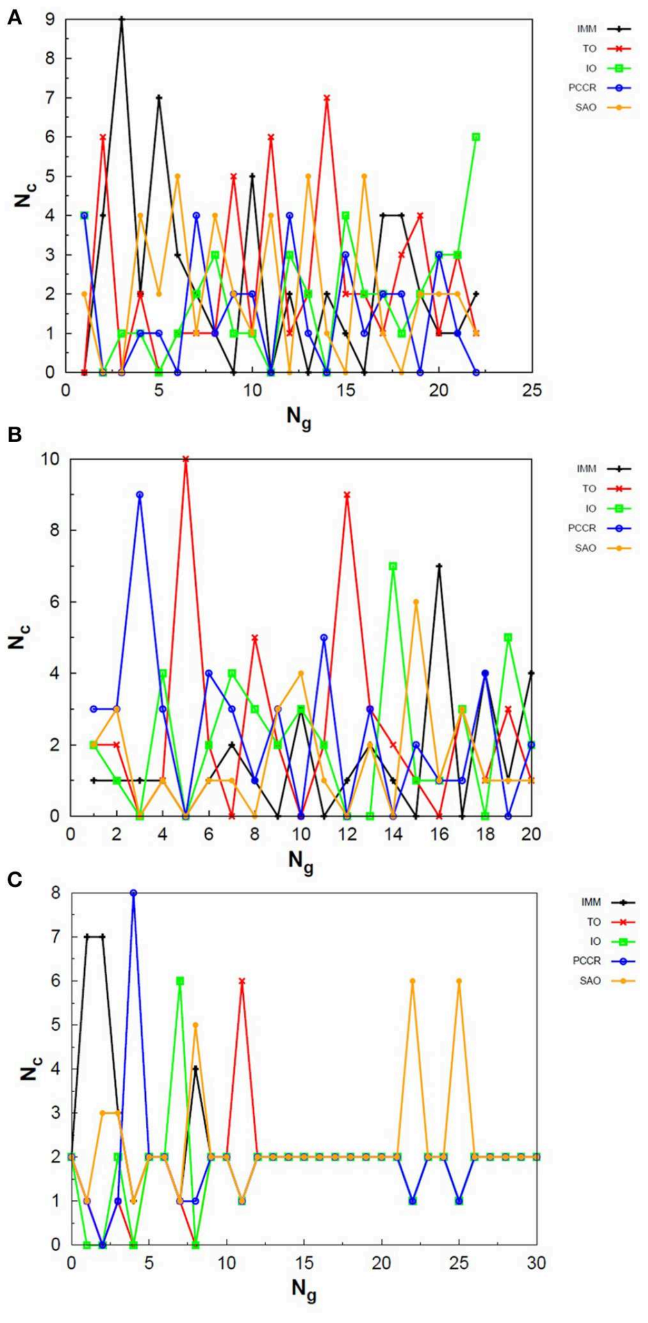


FIGURE 4 | Evolution of the creation of new individuals for AUTO5 build throughout three different runs of the LJ_{26} system, corresponding to the **(A)** third, **(B)** thirty-ninth and **(C)** ninth random number seed employed. N_c is the number of individuals created with that operator and N_g is the generation index. In general, the graphs show large variation in creation rates in the first generations and smaller variations at the end of the simulations.

In general, these creation rates undergo great variations during the first generations and converge to smaller oscillations within a narrower range as the process advances. This can be better noticed when the number of generations needed to reach the global minimum is greater, such as in **Figure 5C**. In the graphs of **Figure 4** one can also notice the importance of TO to the AUTO5 build, which indeed was the responsible for the largest average creation rate in several runs of that build concerning the LJ_{26} system. Still about AUTO5, it is interesting to note that some operators seem to be more important at different

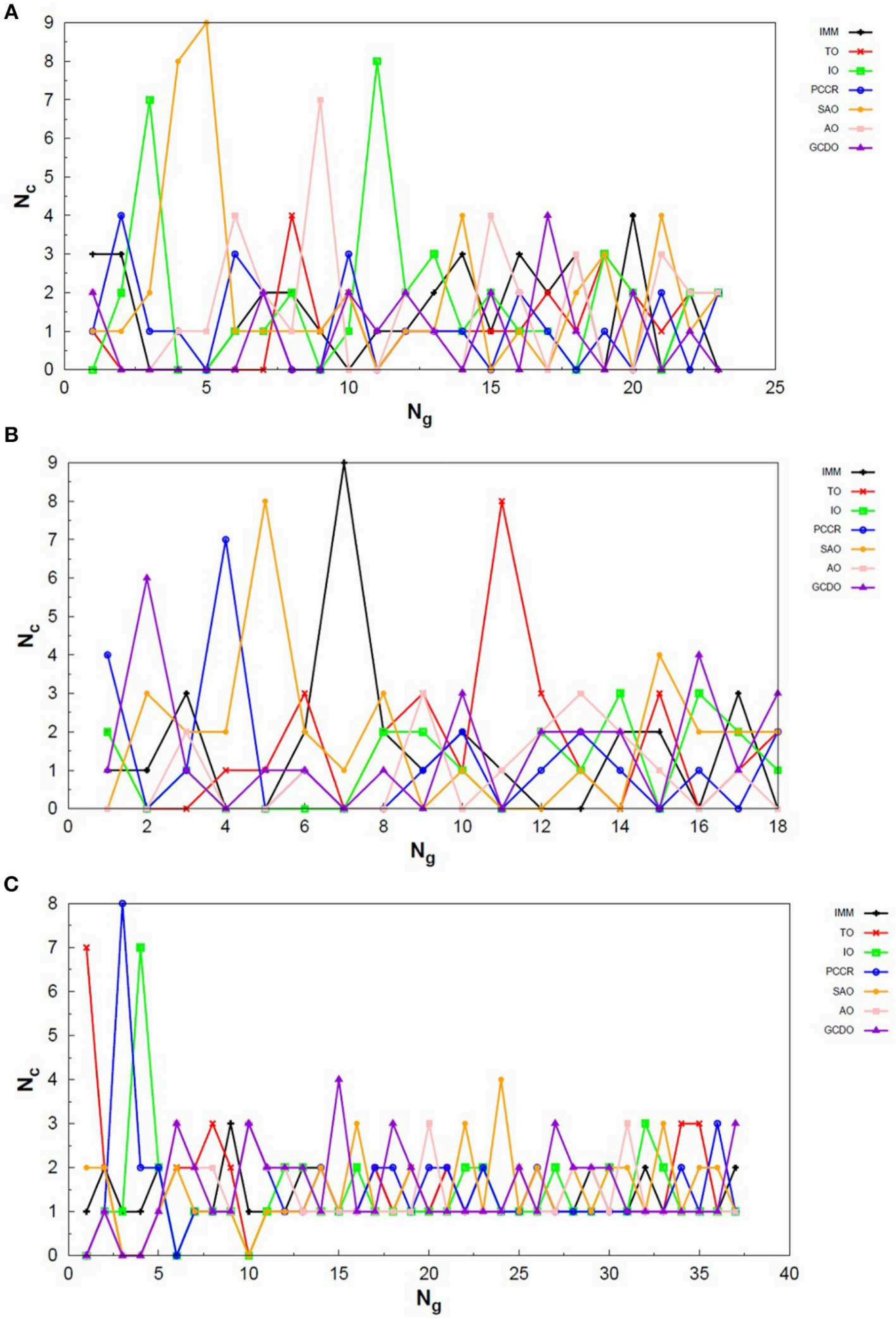


FIGURE 5 | Evolution of the creation of new individuals for AUTO7 build throughout three different runs of the LJ_{26} system, corresponding to the **(A)** thirty-seventh, **(B)** forty-fourth and **(C)** ninth random number seed employed. N_c is the number of individuals created with that operator and N_g is the generation index. In general, the graphs show large variation in creation rates in the first generations and smaller variations at the end of the simulations.

stages of the evolutionary procedure, while others seem to be more systematic. In **Figures 4A,C** it can be seen that the creation rate of IMM, for example, is larger at the beginning and decreases as the system evolves, while essentially the opposite behavior can be observed for IO in **Figures 4A,B**. IMM creates individuals totally randomly, and thus it could be expected to yield better results in a stage where the population is not sufficiently evolved. IO and TO, in turn, were the ones that presented the best performances when evaluated individually within the GA scheme employed here to approach the LJ_{26} system, which is consistent with their behaviors within AUTO5 build.

Among the builds proposed to test our management methodology, AUTO7 seems to be the most balanced one. For the simple LJ_{26} case, for example, its performance has been

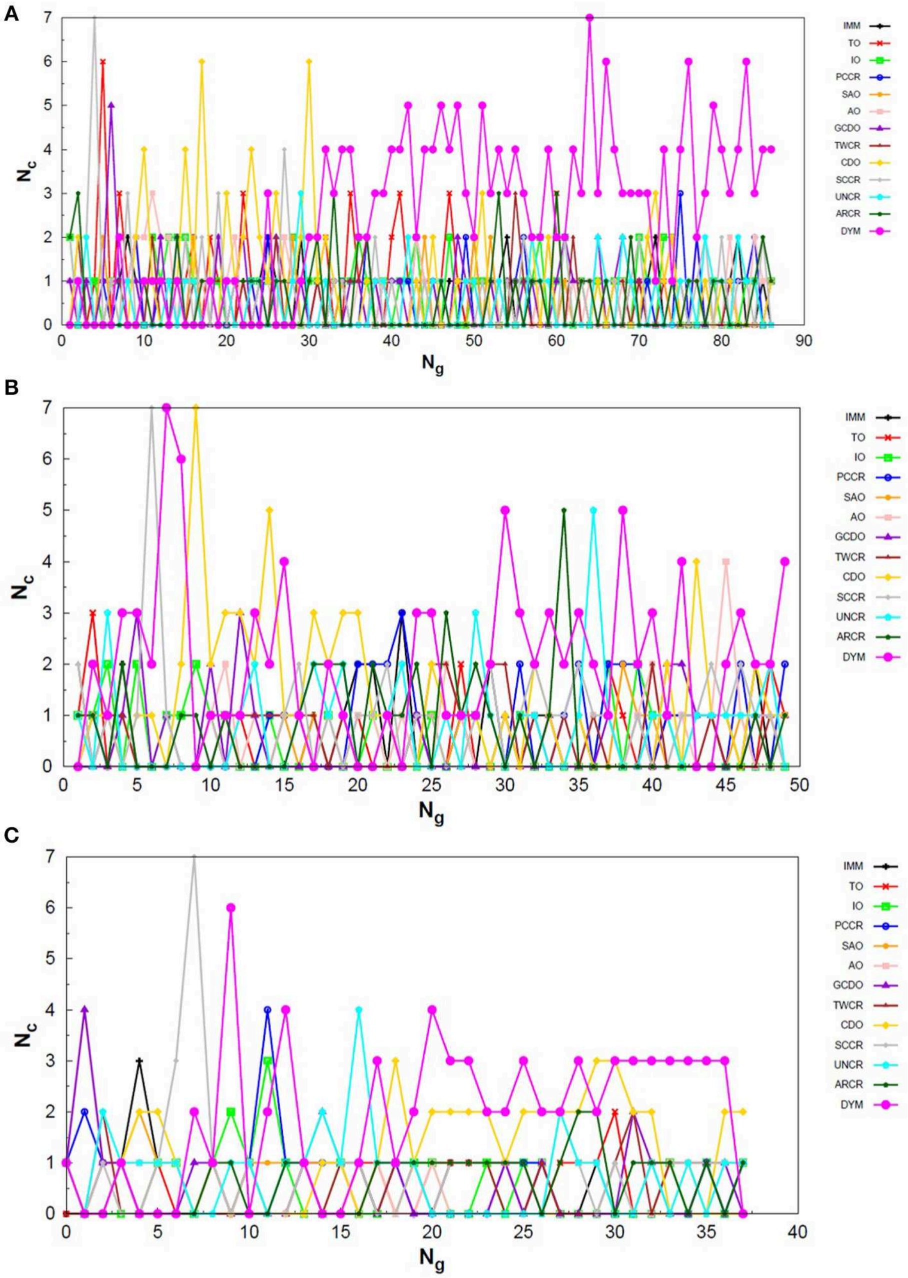


FIGURE 6 | Evolution of the creation of new individuals for AUTO13 build throughout three different runs of the LJ_{26} system, corresponding to the **(A)** seventeenth, **(B)** thirty-ninth and **(C)** ninth random number seed employed. N_c is the number of individuals created with that operator and N_g is the generation index. Except for the peculiar behavior of the DYM operator, the graphs show generally larger variation in creation rates in the first generations and smaller variations at the end of the simulations. The number of generations to reach the global minimum was, on average, greater than that required for the other builds.

essentially equivalent to that of AUTO5, as one can see from $\widehat{N_{LM}}$ in **Table 1** and from the number of generations needed to reach convergence, shown in **Figure 5**. Besides, it has presented, on average, more homogeneous creation rates among its operators throughout the generations when compared to AUTO5 and AUTO13. AUTO7 collects not only the operators that presented the best results when evaluated individually, but also those that failed to converge for some of the random number seeds tested, despite presenting comparable good performance according to $\widehat{N_{LM}}$ (and considering the standard errors). Along with keeping overall performance, this combination allowed a desirable diversity of operator outcomes. Furthermore, this specific combination of operators could enhance the performance of individual ones, such as SAO, which presented the largest

TABLE 2 | Results of tests performed on our GA builds for LJ_{55}.

Build	$\widehat{N_{LM}}$*	$\sigma_{\bar{x}}$	N_{fails}(%)
IO	559	32	0
TO	559	53	0
GCDO	571	41	0
AUTO7	653	37	0
AUTO5	660	46	0
PCCR	775	42	0
AO	830	60	0
SAO	1,380	109	0
AUTO13	864	74	4
TWCR	1,419	190	28
CDO	1,724	223	44

$\widehat{N_{LM}}$ indicates the average number of local minimizations needed to reach the global minimum. $\sigma_{\bar{x}}$ indicates the standard error and N_{fails} indicates the relative number of times the global minimum was not reached.

**The unconverged runs were removed from the calculation of these averages. This removal may compromise the analysis when N_{fails} is nonzero.*

average creation rate in two out of the three runs shown in **Figure 5**.

From the comparison between **Figures 4–6** one can also notice that AUTO13 generally required a considerably larger number of generations to reach global minimum than AUTO7 and AUTO5, which was already expected due to the results shown in **Table 1**. Excluding the DYM operator, the graphs in **Figure 6** also show generally larger variations in creation rates in the first generations and more steady behavior in later generations. However, by analyzing the graphs in **Figure 6**, we can conclude that the presence of operators that performed badly when evaluated individually indeed contributed to the worse performance of AUTO13. Operators such as CDO, UNCR, and DYM managed to create individuals good enough to raise their creation rates, but not sufficiently good to reach the global minimum. These inadequate operators undermined the action of the most suitable ones to perform the original task of finding the lowest energy structure effectively. This means that AUTO13 frequently got stucked in local minima and probably would not be the best choice to tackle a system for which not much information is already available.

A study completely analogous to that presented so far was carried out for the 55-atom case (LJ_{55}) and the results yielded by the builds tested are presented in **Table 2**. Only the builds that successfully converged in more than 50% of the trial runs are presented. For those cases that failed to converge N_{LM} was defined as the maximum allowed number of local minimizations plus one (5001). Again, unconverged runs were not taken into account to obtain $\widehat{N_{LM}}$. The results are presented primarily in ascending order of N_{fails}; secondarily, in ascending order of $\widehat{N_{LM}}$.

The results obtained for LJ_{55} are essentially consistent with those obtained for the LJ_{26} case. IO and TO were again the ones with best performances and, along with GCDO, AUTO7, AUTO5, PCCR, AO, and SAO, make up the builds that managed to find the global minimum on every run. Among the ones proposed to test our management strategy, AUTO7 and AUTO5 performed equivalently again and, once more, outperformed AUTO13. Just as done in the 26-atom case, we have also compared our top ranked builds (IO and TO) with the widely used plane-cut-splice (PCCR) for the 55-atom case using one-tailed p-value approach (Chaubey, 1993). This time, our automated builds (AUTO5 and AUTO7) did not overlap with PCCR when taking their standard errors into account, thus we have also compared AUTO7 (which was essentially equivalent to AUTO5) with PCCR using one-tailed p-value approach. Again, we have previously tested for the normality of the data generated by these builds using the Ryan-Joiner test (Yap and Sim, 2011), and the normality hypothesis was accepted within a significance level of 0.1 without data discard. IO and TO were better than PCCR with a 99% confidence level, while AUTO7 was better than PCCR with a 95% confidence level.

For this larger system, TWCR, CDO, UNCR, ARCR, and DYM performed even worse than they did for the LJ_{26} case, as expected due to the increase in difficulty to find the global minimum as the number of degrees of freedom of the system increases. This time, however, IMM also performed badly and could not converge a significant amount of runs. On the other hand, AO and GCDO did not fail in any run as they did in the previous case. Again, SCCR performed badly, although it was expected to improve when approaching larger systems (Chen et al., 2013).

Although we have separated operators into classes (crossover, mutation, and immigration) in section 3.5, no distinction was made among them when it comes to the number of individuals created by each type in each generation. This was always set on the fly according to our management strategy described in section 3.4 (or kept fixed for the builds with single operators). As a result, mainly mutation type operators presented good performances within our GA approach, both individually and within the builds composed by various operators. Furthermore, operators that act fully in a random way performed generally better than more complex ones which involve, for example, parameterized mathematical expressions or simply parameters to be defined. The latter may be more suitable for less unbiased GA schemes than the one employed here. Excepting PCCR, crossover operators were greatly outperformed, possibly because they need more elaborate methods to select parents for mating to properly yield results. ARCR and SCCR, for instance, did not present high values for $\widehat{N_{LM}}$, but they did present high N_{fails}. This indicates that these operators might be more sensible to the fitness of the selected parents. Finally, regarding the IMM operator, it is reasonable to expect that it would perform worse for larger systems, since the probability of randomly generating good structures would undoubtedly decrease with the number of atoms.

The management strategy proposed in this work proved to be efficient. The performance of AUTO5, for example, approached the average taken over the performance of its individual operators (TO, IO, PCCR, SAO, and IMM) for the LJ_{26} case. This was measured by taking the average value of $\widehat{N_{LM}}$ over the five cited operators, which equals 240, while $\widehat{N_{LM}}$ associated with AUTO5 was 246. For AUTO7 ($\widehat{N_{LM}} = 264$) we have a similar scenario: the

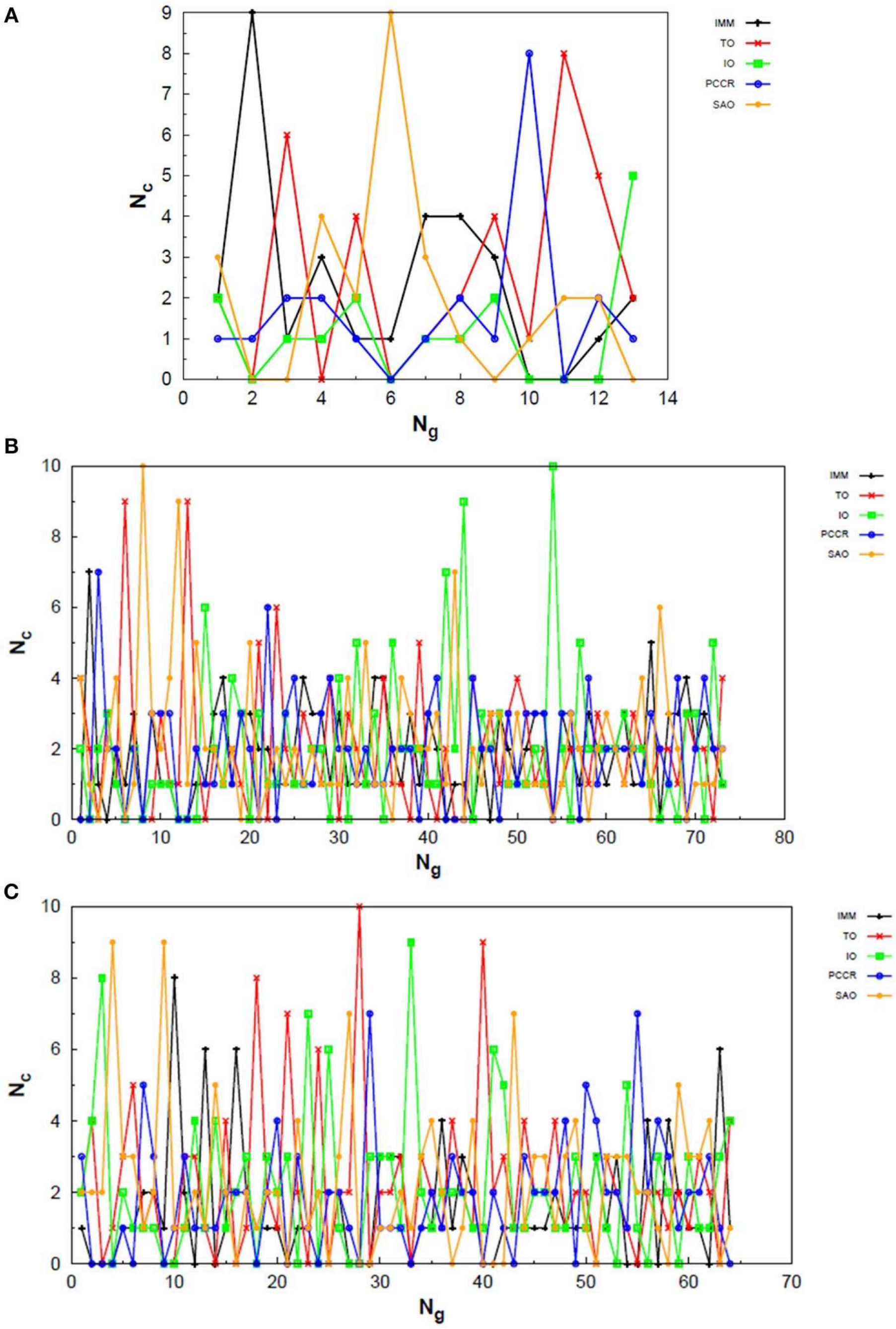


FIGURE 7 | Evolution of the creation of new individuals for AUTO5 build throughout three different runs of the LJ_{55} system, corresponding to the **(A)** seventeenth, **(B)** thirty-ninth and **(C)** forty-fourth random number seed employed. N_c is the number of individuals created with that operator and N_g is the generation index. TO stands out again among the operators of AUTO5 build, followed by SAO and IO.

average of $\widehat{N_{LM}}$ over its individual operators equals 261. On the other hand, AUTO13 ($\widehat{N_{LM}} = 434$) considerably outperformed the average of its operators ($\widehat{N_{LM}} = 620$) and, furthermore, presented only 2% of convergence failure, despite being composed by various operators that showed high failure rate.

For the LJ_{55} case, some operators that presented high values of N_{fails} when employed individually were still added to AUTO5, AUTO7, and AUTO13 builds. In order to compare the performances of these builds with those of their individual operators, however, only the ones shown in **Table 2** were taken into consideration. Thus, for AUTO5 ($\widehat{N_{LM}} = 660$) it was

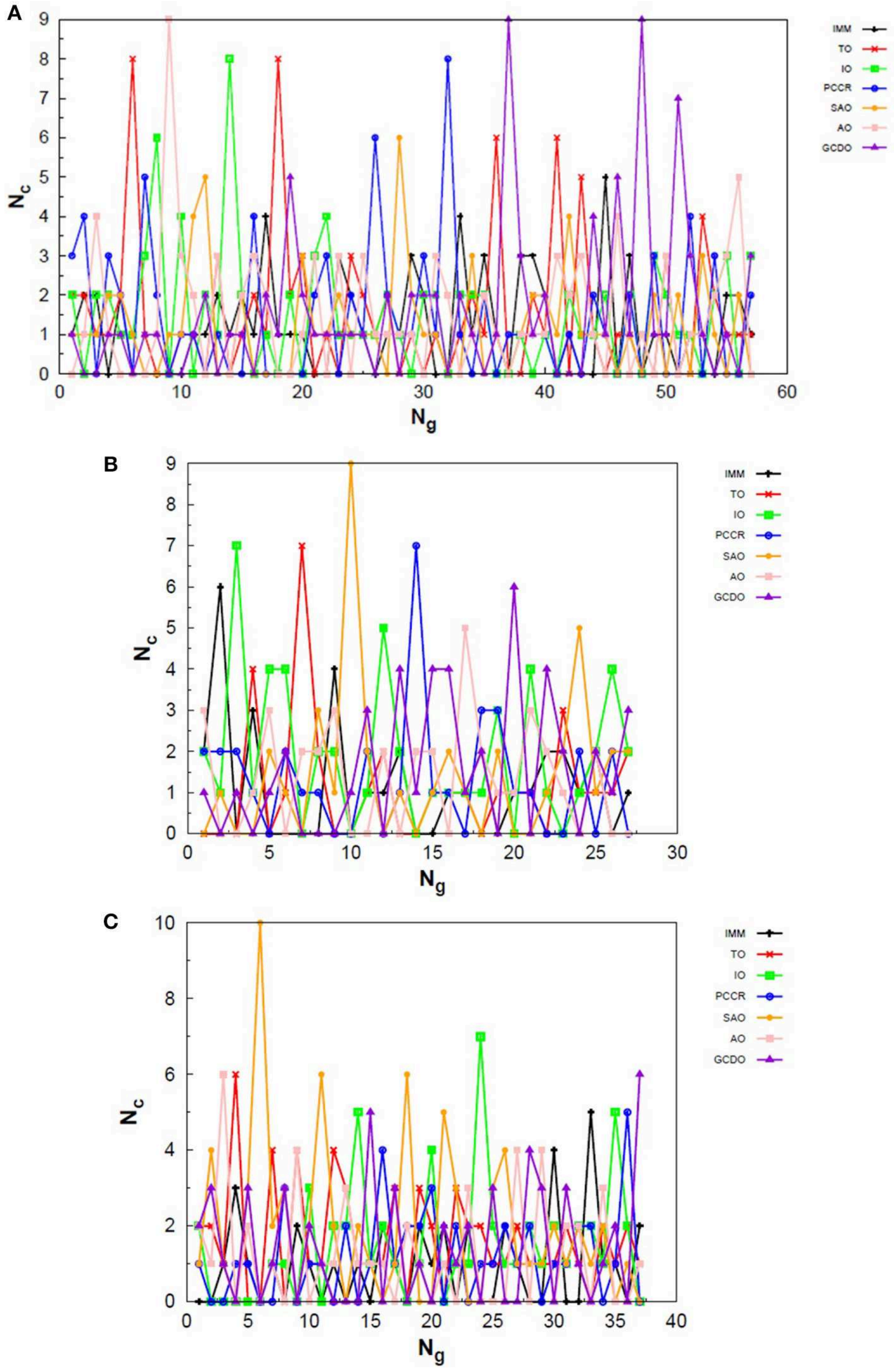


FIGURE 8 | Evolution of the creation of new individuals for AUTO7 build throughout three different runs of the LJ_{55} system, corresponding to the **(A)** third, **(B)** seventeenth and **(C)** thirty-ninth random number seed employed. N_c is the number of individuals created with that operator and N_g is the generation index. AUTO7 was also the most balanced build for the LJ_{55} case, presenting wider diversity of operator outcomes.

measured by taking the average value of $\widehat{N_{LM}}$ over IO, TO, PCCR, and SAO, that equals 818; for AUTO7 ($\widehat{N_{LM}} = 653$), the average value of $\widehat{N_{LM}}$ was taken over IO, TO, GCDO, PCCR, AO, and SAO, which equals 779; for AUTO13 ($\widehat{N_{LM}} = 864$), the average value of $\widehat{N_{LM}}$ was taken over IO, TO, GCDO, PCCR, AO, SAO, TWCR, and CDO, which equals 977. For the larger LJ_{55} cluster, all AUTO builds outperformed the average of their operators. These numbers would favor even further the AUTO builds if the operators omitted from **Table 2** had been taken into account. This time, despite having individually ineffective operators in

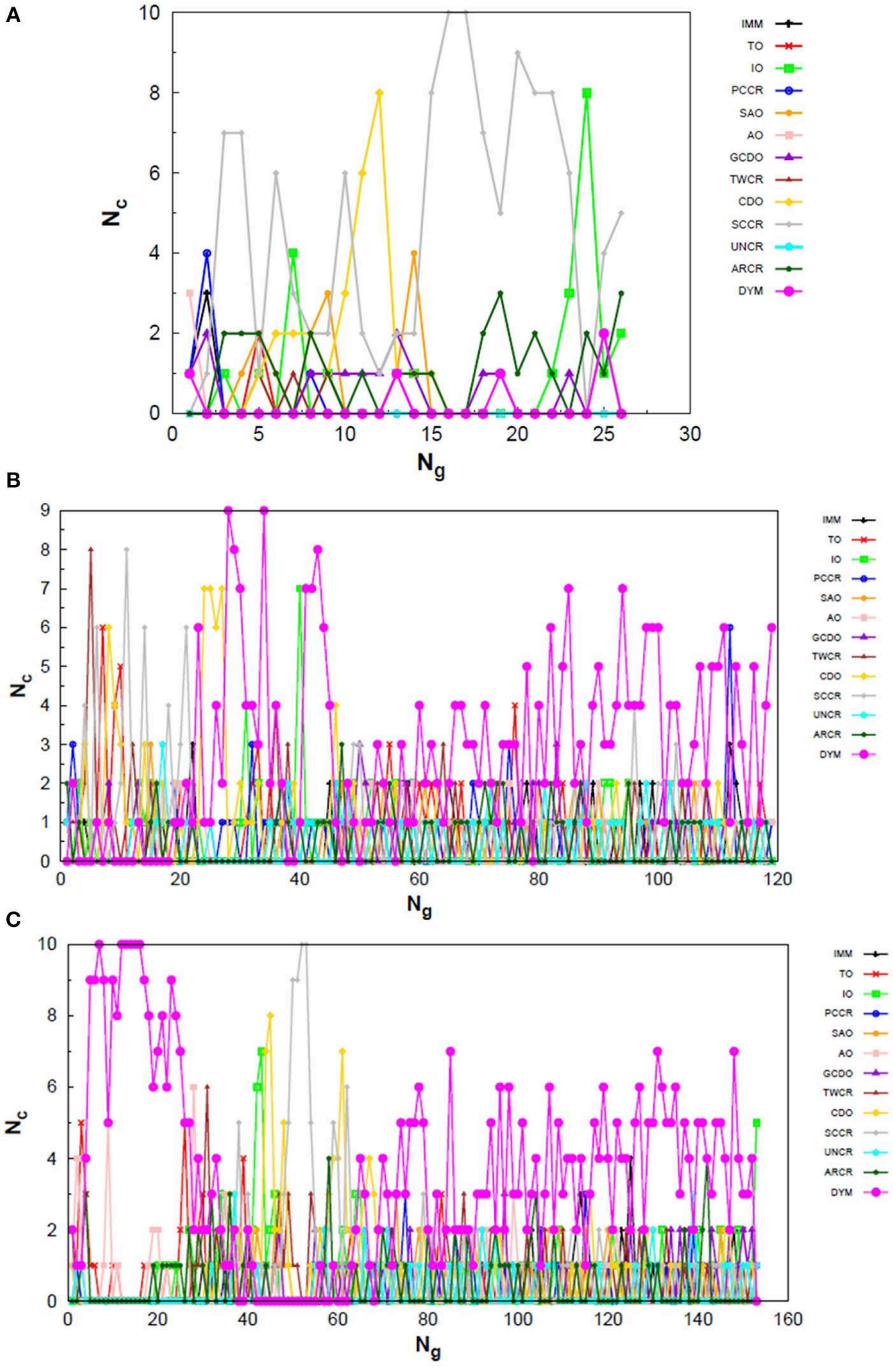


FIGURE 9 | Evolution of the creation of new individuals for AUTO13 build throughout three different runs (randomly chosen) of the LJ_{55} system. N_c is the number of individuals created with that operator and N_g is the generation index. In **(A)** the DYM operator does not act significantly and convergence is reached quickly. In **(B,C)**, again, the DYM operator disturbs the evolutionary procedure causing the number of generations needed to reach the global minimum to be greater than that required for the other builds, on average.

their compositions, all AUTO builds managed to enhance the overall performance.

We have also attempted to perform the same study for LJ_{38}. However, it has a double funnel energy landscape (Chen et al., 2013) and hence is a more complicated system to be approached by our simple GA scheme. Therefore, none of our builds managed to converge to the global energy minimum more than 50% of the 50 trial runs. Builds such as ARCR, UNCR and

TABLE 3 | Average number of local minimizations needed to reach the global minimum ($\widehat{N_{LM}}$) for the C_{18} cluster together with the failure rate (N_{fails}) for each employed build in reaching that minimum.

Build	$\widehat{N_{LM}}$*	N_{fails}(%)
PCCR	1,815	26
AUTO7	1,667	34
AUTO5	2,073	50
AO	1,964	60
AUTO13	1,420	60
GCDO	1,175	62
IO	1,757	66
TO	1,735	86
CDO	1,065	88
TWCR	2,120	92
SCCR	908	94
PREV**	1,505	96
IMM	3,001	100
SAO	3,001	100
ARCR	3,001	100
UNCR	3,001	100
DYM	3,001	100

For all cases, 50 *different random number seeds were executed employing the same parameters of the previous studied systems.*

**The unconverged runs were removed from the calculation of these averages. This removal may compromise the analysis when N_{fails} is nonzero.*

***Previous work (Silva et al., 2015).*

DYM, for example, failed 100% of the trials, while AUTO5 was the best build and managed to find the LJ_{38} global minimum 48% of the times. This is because four out of the five operators that compose AUTO5 build were the ones that presented the highest individual convergence rates. In fact, they were even more effective than AUTO7 and AUTO13. It is interesting to notice, however, that the remaining AUTO5 operator, SAO, failed 94% of the times for the LJ_{38} case. It shows that, despite contaminated by an operator that performed badly individually, AUTO5 managed to outperform every other build when approaching the LJ_{38} system. Once more we have evidence that our management strategy may enhance the overall performance of the method through a synergic action of suitable operators.

Analogously to the LJ_{26} case, we can also assess the variation in the creation rate of the operators within each AUTO build along different GA runs (chosen randomly) regarding the LJ_{55} system. These results are shown in **Figure 7** for AUTO5, **Figure 8** for AUTO7 and **Figure 9** for AUTO13.

For the 55-atom cluster one can still see greater variations in the creation rate of AUTO5 operators up to half of the generations of the runs shown in **Figures 7B,C**. In the same figure (mainly in panels a and c) it can be noticed the same trend observed for the IMM operator when approaching LJ_{26} with AUTO5 build: it has higher creation rates at the beginning and it gets lower as the system evolves. Again, TO was the responsible for the largest average creation rate for this build, which can be inferred from the graphs of **Figure 7**. This is also consistent with the results presented in **Table 2**, where TO appears as the one with best performance when approaching LJ_{55}, along with IO. The interior operator, however, has the third largest average value for the creation rate in this case, being overcome, surprisingly, by SAO. This exemplifies that the combination of operators may enhance their individual performance. From the graphs of **Figure 7** one can also note that SAO influenced mainly the initial stages of the presented runs.

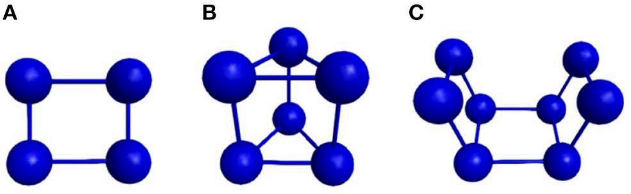


FIGURE 10 | Local energy minima correctly found by QGA-7 for **(A)** N_4, **(B)** N_6, and **(C)** N_8.

AUTO7 was the most balanced build for LJ_{55}, as well as it was for LJ_{26}. It presented more homogeneous distribution of peaks throughout the generations in the graphs of **Figure 8** when compared to the other AUTO builds. None of its operators has been systematically the one with the greatest average creation rate within the evaluated runs. AUTO7 has also required less generations than AUTO5 and AUTO13, on average, to reach convergence, as it can be seen by comparing the graphs in **Figures 7–9**. From the same comparison, we can see that AUTO13 was again the build that generally required more generations to find the global minimum.

Through the analysis of **Figures 9B,C**, it becomes clear that DYM operator disturbed the evolutionary procedure and prevented these runs from converging earlier. In fact, **Figure 9** shows three distinct scenarios: in (a) SCCR dominates the process from the beginning and leaves no room for DYM. Accordingly, the GA converges in only 26 generations. In (b) DYM also starts with low creation rate, but it increases rapidly within a few generations and basically dominates the process from the 24^{th} generation on. The GA converges after 119 generations. In (c) DYM already starts with high creation rate and, although it oscillates and goes through a minimum for approximately 20 generations, it increases again and dominate the process until convergence is reached after 153 generations. As well as it happened to the 26-atom case, the algorithm has spent several generations trapped in local minima while the unsuitable DYM operator prevents other operators from acting and reestablishing the needed population diversity. By all means, it is interesting to notice that AUTO13 managed to find the correct LJ_{55} structure mainly under the influence of operators that failed almost 100% of the times they were tested individually.

In **Table 3** one can find the average number of local minimizations needed to reach the lowest energy structure of C_{18}, as well as the failure rate (N_{fails}) of each build employed here to tackle the C_{18} system within the REBO potential approach. This failure rate indicates the relative number of times the global minimum was not reached by our algorithm. For the C_{18} case, this minimum corresponds to the carbon

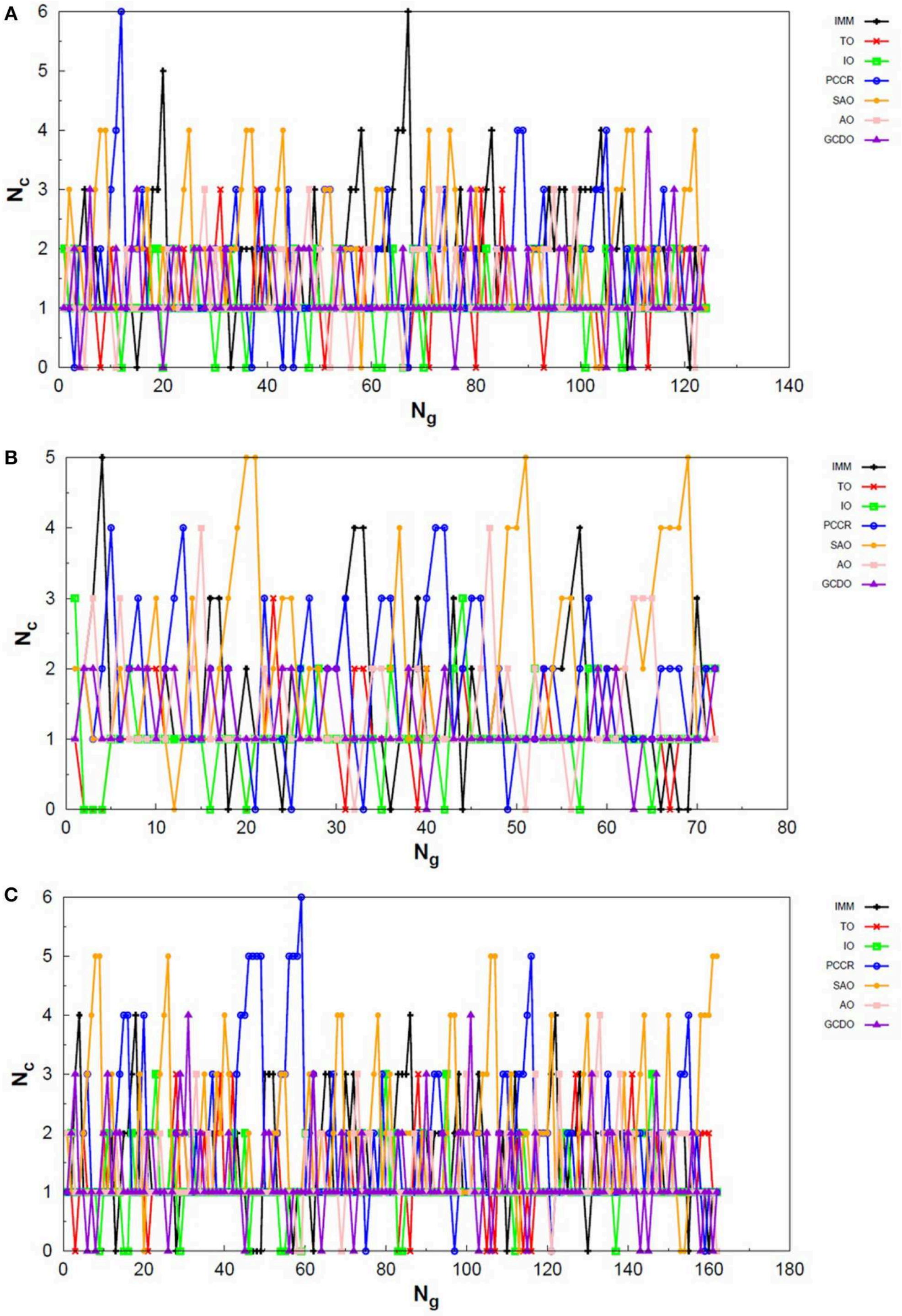


FIGURE 11 | Evolution of the creation of new individuals for QGA-7 throughout three different runs of the N4 system, corresponding to the following random number seeds employed: **(A)** 17, **(B)** 29 and **(C)** 6217. N_c is the number of individuals created with that operator and N_g is the generation index.

atoms arranged in a planar single ring (Kosimov et al., 2010). This cyclocarbon molecule was indeed synthetized by Kaiser et al. using atom manipulation by eliminating carbon monoxide from a cyclocarbon oxide molecule, and characterized by high-resolution atomic force microscopy (Kaiser et al., 2019).

We observed that operators leading geometries toward spherical shape were disadvantaged. TO, for example, which had performed well in the previous cases, performed poorly in the present one. That is because torsions would take atoms off the plane, which is not consistent with the energy minimum

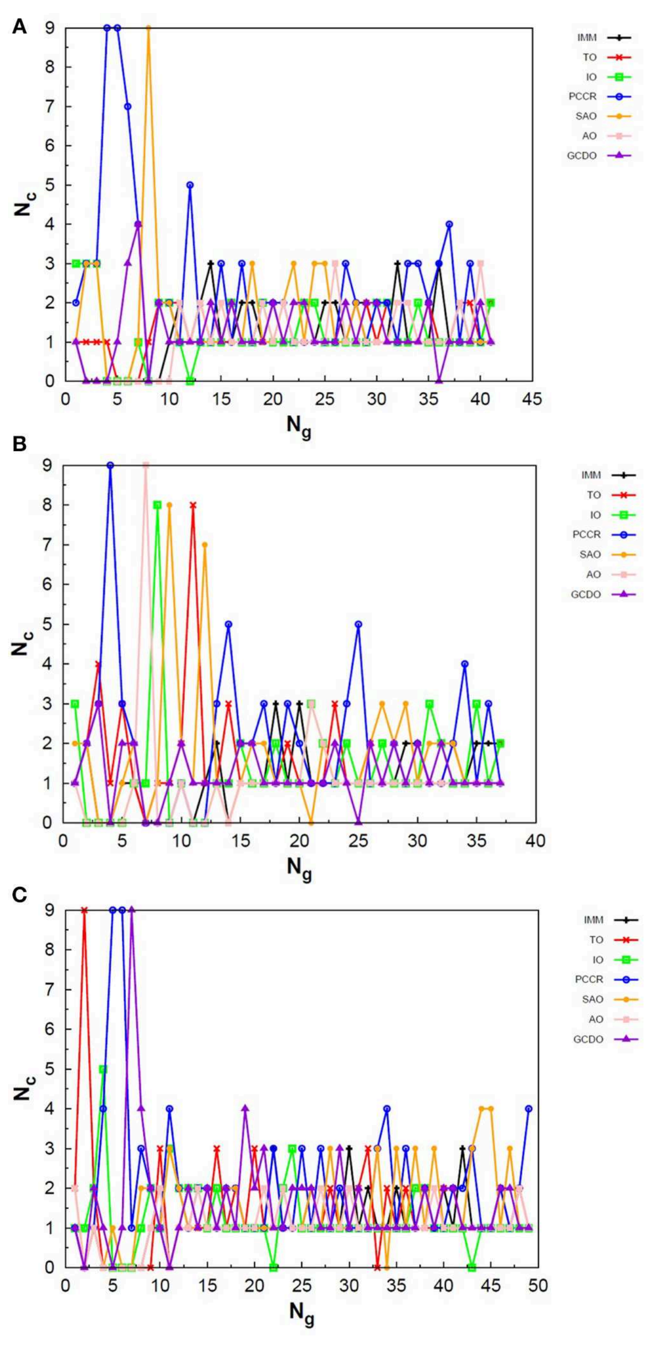


FIGURE 12 | Evolution of the creation of new individuals for QGA-7 throughout three different runs of the N_6 system, corresponding to the following random number seeds employed: **(A)** 9, **(B)** 29 and **(C)** 6217. N_c is the number of individuals created with that operator and N_g is the generation index.

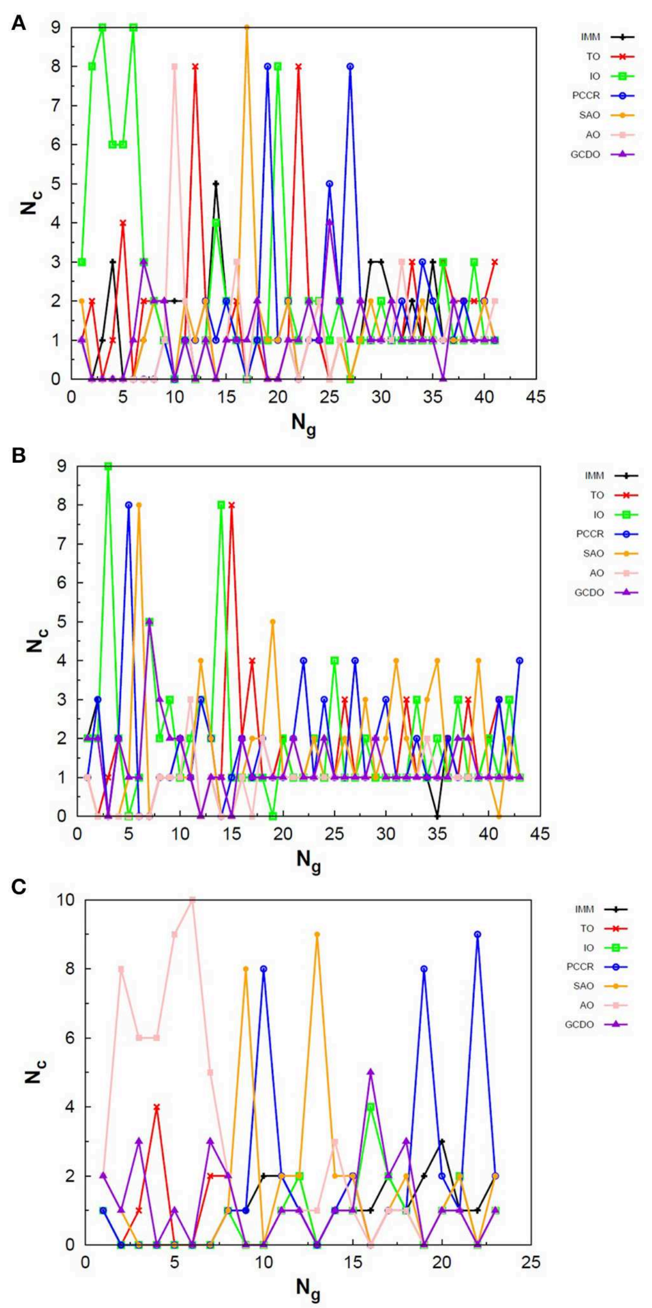


FIGURE 13 | Evolution of the creation of new individuals for QGA-7 throughout three different runs of the N_8 system, corresponding to the following random number seeds employed: **(A)** 29, **(B)** 62 and **(C)** 666. N_c is the number of individuals created with that operator and N_g is the generation index.

for the present system, which is planar. IMM, for instance, which generates atoms randomly inside a sphere, could not reach the global minimum in any run, as one could expect. SAO operator is also biased to generate spherical structures, and could not find the planar energy minimum in any run. Analogous argument can be used for the poor performance of SCCR and PREV, for example. On the other hand, the best results for the C_{18} case were obtained by the PCCR build, probably due to the use of planes to slice each parent cluster in crossover step.

The builds proposed in the present work (AUTO5, AUTO7, and AUTO13) presented essentially the same behavior observed in the previous cases, in which the stability of results is maintained, benefiting the best operators and avoiding the worst ones for each specific case. Our management strategy provides the useful advantage of versatility to the optimization algorithm. Despite the best performances for the cases tested here have been obtained by builds composed by individual operators, the AUTO builds can make the algorithm more adaptable to a wider range of problems. Therefore, we believe that our management strategy would be useful to improve the exploration of PES associated with more complex systems when applied together with a more robust GA framework, which consists the next step of our research.

Based on the analysis carried out so far, we have chosen AUTO7 to apply our strategy within a quantum approach. In order to do so, we have incorporated this build to a more robust GA scheme, namely QGA, which was also coupled to GAMESS-US (Schmidt et al., 1993) and adapted to approach dissociative systems. We have already used QGA to approach polynitrogen systems and to predict good candidates for high energy density materials (HEDM) (Silva et al., 2018), and now we intend to run QGA with AUTO7 (referred to as QGA-7 from now on) to reproduce some energy minima found in our previous work and to evaluate the behavior of the operators along the generations. These polynitrogens are atomic nitrogen clusters, which means that they form structures with nitrogen atoms connected to each other mainly by single or double bonds. Therefore, these clusters consist of local energy minima on the PES, while the global minimum consists of the dissociated system into N_2 molecules.

In **Figure 10** one can find the structures corresponding to local energy minima on the PES associated with N_4, N_6 and N_8 that we managed to reproduce with QGA-7 within a DFT approach employing B3LYP exchange and correlation functional and 6-31G basis set. Besides that, the variation in the creation rate of the operators within QGA-7 along different GA runs (chosen randomly) regarding the N_4 system is shown in **Figure 11**. The same analysis concerning the N_6 and N_8 systems are presented in **Figures 12**, **13**, respectively.

The graphs in **Figure 11** show almost periodic oscillations in the creation rate of each operator involved in QGA-7. These oscillations have essentially constant amplitude for each operator over the entire evolutionary procedure, and are also quite homogeneous among the different ones. This, along with the large number of generations needed to reach convergence for such a small system, indicates stagnation. This can be explained by the fact that the D_{2h} structure (**Figure 10A**) for tetranitrogen is not that far from the much more stable $2N_2$ system within a random structure generator scheme perspective. Since QGA-7 was prepared so as not to allow these N_2 fragments to get too far apart from each other, several quasi-degenerate structures may be generated before the optimum distance between these two moieties is reached. Nevertheless, QGA-7 still converged within the criterium established (an individual remained as the one with the lowest energy for 20 consecutive generations), and the operators that stood out, on average, were SAO, PCCR and IMM.

Differently from the behavior presented by the N_4 system, the creation rate of the operators for the N_6 and N_8 cases resembled that observed for the LJ_{26} system, with greater variations along the first generations which stabilize to become smaller oscillations until convergence is reached. In fact, QGA-7 found the correct structures much more efficiently for these cases than for tetranitrogen. If we do not take into consideration the noisy initial part of the evolutionary procedure, it can be seen that, in general, the operators that stood out were PCCR and SAO. It is interesting to notice that SAO had an important role both in the Lennard-Jones and in the quantum approach of atomic clusters, while PCCR stood out mainly within the quantum approach and TO and IO stood out mainly within the classical approach.

Although we performed only a few simple tests with QGA-7, our management strategy applied to a more complex GA scheme and within a quantum approach was consistent with our primary tests on Lennard-Jones clusters. The results obtained so far may guide us toward the next steps to improve our algorithms, incorporate more efficient builds and enhance its performance to approach more complex systems.

Some well-identified problematic cases [such as LJ_{38}, LJ_{75-77}, LJ_{98}, $LJ_{102-104}$ and some short-ranged Morse clusters (Hartke, 1999; Cheng et al., 2004; Pereira and Marques, 2009)] must still be properly addressed in order to ensure that our strategy is indeed effective in exploring potential energy surfaces in a more extensive way. To do so, it is interesting that our algorithm become independent of extra information about the problem and less system-specific, that is more versatile, while maintaining population diversity (Lee et al., 2003; Grosso et al., 2007). We are currently implementing a new step in which all structures involved in each generation will be compared to each other so that we can evaluate structure similarities and avoid population stagnation. Within this future approach, even enantiomers may be told apart, and population diversity will be greatly enhanced. Different rules for the variation of the application rate of operators will also be tested, and not only the energy of the offspring may be compared to the average energy of the previous population, but also the capability of the applied operator to generate diverse structures. Furthermore, employing a less deterministic selection step, together with a combination of crossover and mutation operators to generate descendants may be also essential to help our algorithm tackle these harder optimization scenarios.

Nevertheless, the management strategy proposed in this work has already proved to be quite promising. Despite some single operator builds have performed better than the management methodologies tested (AUTO) for the LJ_{26} and LJ_{55} cases, this may not hold for larger and more complex systems, as well as for *ab initio* or DFT-based genetic algorithms. We propose that greater versatility might be essential to efficiently sample the PES and to avoid stagnating into a population with serious (SCF or structure optimization) convergence problems, mainly in the first generations.

5. CONCLUSIONS

We have developed a method that manages the creation rate of evolutionary operators within a genetic algorithm procedure on the fly. Its performance was evaluated on 26 and 55-atom Lennard-Jones clusters (LJ_{26} and LJ_{55}) and the obtained results show that our strategy proved to be quite efficient. Moreover, we have assessed thirteen operators available in the literature and, within our simple and highly ubiased GA approach, twist operator was faster than commonly used Deaven and Ho's plane-cut-splice crossover. Also, interior and surface operators, formerly designed for basin-hopping methodology (Takeuchi, 2007; Ye et al., 2011), performed well in our genetic algorithm scheme, although they may have been favored due to the essentially spherical shape of the global energy minima approached.

Three different builds were proposed to test our management strategy, namely AUTO5, AUTO7, and AUTO13, where the numbers indicate how many creation operators were employed in each build. For the LJ_{26} case, the performances of AUTO5 and AUTO7 approached the average taken over the performances of their individual operators. This was measured by taking the average value of $\widehat{N_{LM}}$ (average number of local energy minimizations) over the cited individual operators. On the other hand, AUTO13 considerably outperformed the average of its operators and, furthermore, presented only 2% of convergence failure, despite being composed by various operators that showed high failure rate.

For the LJ_{55} case, some operators that presented high failure rates when employed individually were still added to AUTO5, AUTO7, and AUTO13 builds. However, in order to compare the performances of these builds with those of their individual operators, only the latter ones that successfully converged in more than 50% of the trial runs were taken into consideration. Following this protocol, all AUTO builds outperformed the average of their individual operators. The numbers would favor even further the AUTO builds if all the individual operators tested had been taken into account, regardless of their failure rates. This time, despite having individually ineffective operators in their compositions, all AUTO builds managed to enhance the overall performance.

When tackling the C_{18} system, which presents a planar ring structure as the lowest energy minimum, we could observe that operators that relied mainly on spherical-based creation or transformations of individuals performed poorly, as one could expect. Our management strategy benefited the most appropriate operators and avoided the worst ones, making the algorithm more adaptable and versatile.

These results indicate that our management strategy could benefit from the advantages of the employed operators without loosing overall performance. It may actually enhance the overall performance and help to better explore the parameter space through the diverse combinations of appropriate evolutionary operators and efficient genetic algorithm schemes.

When approaching systems where the global minimum is not known, it is generally hard to tell which operator is the most suitable or efficient to promote GA convergence. Thus, employing several techniques combined and properly managing their application throughout the evolutionary procedure could be the best approach. Among the proposed builds, AUTO7, which combines diversity with speed, was the one chosen to be incorporated in a more robust GA scheme to test our strategy within a quantum approach to polynitrogen systems. This application of our management strategy was consistent with our simpler approach involving Lennard-Jones clusters. We have also managed to find correct polynitrogen structures and to evaluate the behavior of the creation rate of the operators involved in the proposed build within the quantum approach.

With the results yielded by this study we may be able to improve our builds by combining more appropriate operators, as well as our genetic algorithm itself, by implementing more efficient steps that could lead to faster convergence. This would be useful for further cluster studies, which may include *ab initio* and DFT potential energy surface survey.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/supplementary material.

AUTHOR CONTRIBUTIONS

FS: writing the code for the presented operator management strategy, including the employed genetic algorithm itself, as well as reviewing the literature for choosing the operators implemented in the algorithm and writing the paper. MS: running calculations to test the algorithm and the management strategy within the chosen Lennard-Jones systems, compiling, interpreting and evaluating the results, proposing needed modifications both to the employed strategy and to the final algorithm, taking the study to a quantum approach to polynitrogen systems and writing the paper. JB: guiding the studies, revising the paper and managing laboratory resources and research funding.

FUNDING

This work has the financial support from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do estado de Minas Gerais (FAPEMIG) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

REFERENCES

Bader, S. D. (2006). Colloquium: opportunities in nanomagnetism. *Rev. Mod. Phys.* 78, 1–15. doi: 10.1103/RevModPhys.78.1

Bonnin, M. A., Falvo, C., Calvo, F., Pino, T., and Parneix, P. (2019). Simulating the structural diversity of carbon clusters across the planar-to-fullerene transition. *Phys. Rev. A* 99:042504. doi: 10.1103/PhysRevA.99.042504

Borguesan, B., Silva, M. B., Grisci, B., Ponta, M. I., and Dorn, M. (2015). Apl: an angle probability list to improve knowledge-based metaheuristics for the three-dimensional protein structure prediction. *Comput. Biol. Chem.* 59, 142–157. doi: 10.1016/j.compbiolchem.2015.08.006

Brenner, D. W., Shenderova, O. A., Harrison, J. A., Stuart, S. J., Ni, B., and Sinnott, S. B. (2002). A second-generation reactive empirical bond order (rebo) potential energy expression for hydrocarbons. *J. Phys.* 14, 783–802. doi: 10.1088/0953-8984/14/4/312

Burton, A. R., and Vladimirova, T. (1998). "Genetic algorithm utilising neural network fitness evaluation for musical composition," in *Artificial Neural Nets and Genetic Algorithms*, eds G. D. Smith, N. C. Steele, and R. F. Albrecht (Vienna: Springer), 219–223.

Cai, W., Feng, Y., Shao, X., and Pan, Z. (2002). Optimization of lennard-jones atomic clusters. *J. Mol. Struct. THEOCHEM* 579, 229–234. doi: 10.1016/S0166-1280(01)00730-8

Chaubey, Y. P. (1993). Resampling-based multiple testing: examples and methods for p-value adjustment. *Technometrics* 35, 450–451.

Chen, Z., Jiang, X., Li, J., and Li, S. (2013). A sphere-cut-splice crossover for the evolution of cluster structures. *J. Chem. Phys.* 138:214303. doi: 10.1063/1.4807091

Cheng, L., Cai, W., and Shao, X. (2004). A connectivity table for cluster similarity checking in the evolutionary optimization method. *Chem. Phys. Lett.* 389, 309–314. doi: 10.1016/j.cplett.2004.03.125

Deaven, D. M., and Ho, K. M. (1995). Molecular geometry optimization with a genetic algorithm. *Phys. Rev. Lett.* 75, 288–291. doi: 10.1103/PhysRevLett.75.288

Froltsov, V. A., and Reuter, K. (2009). Robustness of 'cut and splice' genetic algorithms in the structural optimization of atomic clusters. *Chem. Phys. Lett.* 473, 363–366. doi: 10.1016/j.cplett.2009.04.015

Götz, D. A., Heiles, S., Johnston, R. L., and Schäfer, R. (2012). Note: gas phase structures of bare si8 and si11 clusters from molecular beam electric deflection experiments. *J. Chem. Phys.* 136:186101. doi: 10.1063/1.4717708

Grosso, A., Locatelli, M., and Schoen, F. (2007). A population-based approach for hard global optimization problems based on dissimilarity measures. *Math. Program. Ser. A* 110, 373–404. doi: 10.1007/s10107-006-0006-3

Guimarães, F. F., Belchior, J. C., Johnston, R. L., and Roberts, C. (2002). Global optimization analysis of water clusters (h2o)n $(11 \leq n \leq 13)$ through a genetic algorithm evolutionary approach. *J. Chem. Phys.* 116, 8327–8333. doi: 10.1063/1.1471240

Hartke, B. (1999). Global cluster geometry optimization by a phenotype algorithm with niches: location of elusive minima, and low-order scaling with cluster size. *J. Comput. Chem.* 20, 1752–1759.

Heiles, S., and Johnston, R. L. (2013). Global optimization of clusters using electronic structure methods. *Int. J. Quantum Chem.* 113, 2091–2109. doi: 10.1002/qua.24462

Heiles, S., Logsdail, A. J., Schäfer, R., and Johnston, R. L. (2012). Dopant-induced 2d-3d transition in small au-containing clusters: dft-global optimisation of 8-atom au-ag nanoalloys. *Nanoscale* 4, 1109–1115. doi: 10.1039/C1NR11053E

Islas, R., Heine, T., Ito, K., Schleyer, P. v. R., and Merino, G. (2007). Boron rings enclosing planar hypercoordinate group 14 elements. *J. Am. Chem. Soc.* 129, 14767–14774. doi: 10.1021/ja074956m

Jiang, H., Kammler, M., Ding, F., Dorenkamp, Y., Manby, F. R., Wodtke, A. M., et al. (2019). Imaging covalent bond formation by h atom scattering from graphene. *Science* 364, 379–382. doi: 10.1126/science.aaw6378

Jiménez-Halla, J. O. C., Islas, R., Heine, T., and Merino, G. (2010). B19-: An aromatic wankel motor," *Angew. Chem. Ind. Ed.*, vol. 49, pp. 5668–5671.

Jin, Y., Olhofer, M., and Sendhoff, B. (2002). A framework for evolutionary optimization with approximate fitness functions. *IEEE Trans. Evolut. Comput.* 6, 481–494. doi: 10.1109/TEVC.2002.800884

Johnston, R. L. (2003). Evolving better nanoparticles: genetic algorithms for optimising cluster geometries. *Dalton Trans.* 22, 4193–4207. doi: 10.1039/B305686D

Jones, J. E., and Ingham, A. E. (1925). On the calculation of certain crystal potential constants, and on the cubic crystal of least potential energy. *Proc. R. Soc. Lond. A* 107, 636–653. doi: 10.1098/rspa.1925.0047

Kaiser, K., Scriven, L. M., Schulz, F., Gawel, P., Gross, L., and Anderson, H. L. (2019). An sp-hybridized molecular carbon allotrope, cyclo[18]carbon. *Science* 365, 1299–1301. doi: 10.1126/science.aay1914

Kim, H. G., Choi, S. K., and Lee, H. M. (2008). New algorithm in the basin hopping monte carlo to find the global minimum structure of unary and binary metallic nanoclusters. *J. Chem. Phys.* 128:144702. doi: 10.1063/1.2900644

King, D. E. (2009). Dlib-ml: a machine learning toolkit. *J. Mach. Learn. Res.* 10, 1755–1758. doi: 10.1145/1577069.1755843

Kosimov, D. P., Dzhurakhalov, A. A., and Peeters, F. M. (2010). Carbon clusters: from ring structures to nanographene. *Phys. Rev. B*, 81:195414. doi: 10.1103/PhysRevB.81.195414

Larsen, A. H., Mortensen, J. J., Blomqvist, J., Castelli, I. E., Christensen, R., Dułak, M., et al. (2017). The atomic simulation environment—a python library for working with atoms. *J. Phys. Condens. Matter* 29:273002. doi: 10.1088/1361-648X/aa680e

Lazauskas, T., Sokol, A. A., and Woodley, S. M. (2017). An efficient genetic algorithm for structure prediction at the nanoscale. *Nanoscale* 9, 3850–3864. doi: 10.1039/C6NR09072A

Lee, J., Lee, I. H., and Lee, J. (2003). Unbiased global optimization of lennard-jones clusters for n < or = 201 using the conformational space annealing method. *Phys. Rev. Lett.* 91:080201. doi: 10.1103/PhysRevLett.91.080201

Lin, X., Zhang, H., Guo, Z., and Chang, T. (2019). Strain engineering of friction between graphene layers. *Tribol. Int.* 131, 686–693. doi: 10.1016/j.triboint.2018.11.028

Lordeiro, R. A., Guimarães, F. F., Belchior, J. C., and Johnston, R. L. (2003). Determination of main structural compositions of nanoalloy clusters of cuxauy $(x + y \leq 30)$ using a genetic algorithm approach. *Int. J. Quantum Chem.* 95, 112–125. doi: 10.1002/qua.10660

Louis, S. J., and McDonnel, J. (2004). Learning with case-injected genetic algorithms. *IEEE Trans. Evol. Comput.* 8, 316–328. doi: 10.1109/TEVC.2004.823466

Lu, Y., Xu, Y. J., Zhang, G. B., Ling, D., Wang, M. Q., Zhou, Y., et al. (2017). Iron oxide nanoclusters for t1 magnetic resonance imaging of non-human primates. *Nat. Biomed. Eng.* 1, 637–643. doi: 10.1038/s41551-017-0116-7

Marques, J. M. C., Jesus, W. S., Prudente, F. V., Pereira, F. B., and Lourenço, N. (2018). *Physical Chemistry for Chemists and Chemical Engineers*. New York, NY: Apple Academic Press.

Michalewicz, Z. (1996). *Genetic Algorithms + Data Structures = Evolutionary Programs*. Berlin; Heidelberg: Springer.

Morse, P. M. (1929). Diatomic molecules according to the wave mechanics. II. vibrational levels. *Phys. Rev.* 34, 57–64. doi: 10.1103/PhysRev.34.57

Moseler, M., Häkkinen, H., Barnett, R. N., and Landman, U. (2001). Structure and magnetism of neutral and anionic palladium clusters. *Phys. Rev. Lett.* 86, 2545–2548. doi: 10.1103/PhysRevLett.86.2545

Pelegrini, M., Parreira, R. L. T., Ferrão, L. F. A., Caramori, G. F., Ortolan, A. O., Silva, E. H., et al. (2016). Hydrazine decomposition on a small platinum cluster: the role of n2h5 intermediate. *Theor. Chem. Acc.* 135:58. doi: 10.1007/s00214-016-1816-x

Pereira, F. B., and Marques, J. M. C. (2009). A study on diversity for cluster geometry optimization. *Evol. Intel.* 2, 121–140. doi: 10.1007/s12065-009-0020-5

Rieth, M., and Schommers, W. (2002). Computational engineering of metallic nanostructures and nanomachines. *J. Nanosci. Nanotech.* 2, 679–685. doi: 10.1166/jnn.2002.145

Rodrigues, D. D. C., Nascimento, A. M., Duarte, H. A., and Belchior, J. C. (2008). Global optimization analysis of cunaum $(n + m = 38)$ clusters: complementary ab initio calculations. *Chem. Phys.* 349, 91–97. doi: 10.1016/j.chemphys.2008.02.069

Rondina, G. G., and Da Silva, J. L. (2013). Revised basin-hopping monte carlo algorithm for structure optimization of clusters and nanoparticles. *J. Chem. Inf. Model.* 53, 2282–2298. doi: 10.1021/ci400224z

Saini, N. (2017). Review of selection methods in genetic algorithms. *Int. J. Eng. Comput. Sci.* 6, 22261–22263. doi: 10.18535/ijecs/v6i12.04

Schmidt, M. W., Baldridge, K. K., Boats, J. A., Elbert, S. T., Gorgon, M. S., Jensen, J. H., et al. (1993). General atomic and molecular electronic structure system. *J. Comput. Chem.* 14, 1347–1363. doi: 10.1002/jcc.5401 41112

Silva, F. T., Galvão, B. R. L., Voga, G. P., Silva, M. X., Rodrigues, D. D. C., and Belchior, J. C. (2015). Exploring the mp2 energy surface of nanoalloy clusters with a genetic algorithm: application to sodium-potassium. *Chem. Phys. Lett.* 639, 135–141. doi: 10.1016/j.cplett.2015.09.016

Silva, M. X., Galvão, B. R., and Belchior, J. C. (2014a). Growth analysis of sodium-potassium alloy clusters from 7 to 55 atoms through a genetic algorithm approach. *J. Mol. Model.* 20:2421. doi: 10.1007/s00894-014-2421-3

Silva, M. X., Galvão, B. R. L., and Belchior, J. C. (2014b). Theoretical study of small sodium-potassium alloy clusters through genetic algorithm and quantum chemical calculations. *Phys. Chem. Chem. Phys.* 16, 8895–8904. doi: 10.1039/C3CP55379E

Silva, M. X., Silva, F. T., Galvão, B. R. L., and Belchior, J. C. (2018). A genetic algorithm survey on closed-shell atomic nitrogen clusters employing a quantum chemical approach. *J. Mol. Model.* 24:196. doi: 10.1007/s00894-018-3724-6

Song, S., Gao, S., Chen, X., Jia, D., Qian, X., and Todo, Y. (2018). Aimoes: archive information assisted multi-objective evolutionary strategy for ab initio protein structure prediction. *Knowl. Based Syst.* 146, 58–72. doi: 10.1016/j.knosys.2018.01.028

Takeuchi, H. (2007). Novel method for geometry optimization of molecular clusters: application to benzene clusters. *J. Chem. Inf. Model.* 47, 104–109. doi: 10.1021/ci600336p

Vilhelmsen, L. B., and Hammer, B. (2012). Systematic study of au6 to au12 gold clusters on mgo(100) f centers using density-functional theory. *Phys. Rev. Lett.* 108:126101. doi: 10.1103/PhysRevLett.108.126101

Wales, D. J., and Doye, J. P. K. (1997). Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms. *J. Phys. Chem. A* 101, 5111–5116. doi: 10.1021/jp970984n

Wang, J., Du, N., and Chen, H. (2018). Structure and stability of al n mg m (n = 4 8, m = 1 3) clusters: genetic algorithm and density functional theory approach. *Comput. Theor. Chem.* 1128, 15–23. doi: 10.1016/j.comptc.2018.02.006

Yan, X., and Wang, X. (2010). "Fitness function of genetic algorithm in structural constraint optimization," in *Advances in Swarm Intelligence. Lecture Notes in Computer Science, Vol. 6145*, eds Y. Tan, Y. Shi, and K. C. Tan (Berlin; Heidelberg: Springer), 432–438.

Yap, B. W., and Sim, C. H. (2011). Comparison of various types of normality tests. *J. Stat. Comput. Sim.* 81, 2141–2155. doi: 10.1080/00949655.2010.520163

Ye, T., Xu, R., and Huang, W. (2011). Global optimization of binary lennard-jones clusters using three perturbation operators. *J. Chem. Inf. Model.* 51, 572–577. doi: 10.1021/ci1004256

Zayed, A. O. H., Daud, M. N., and Zain, S. M. (2017). Global structural optimization and growth mechanism of cobalt oxide nanoclusters by genetic algorithm with spin-polarized dft. *J. Alloys Compd.* 695, 2513–2518. doi: 10.1016/j.jallcom.2016.11.153

Zhao, J., Shi, R., Sai, L., Huang, X., and Su, Y. (2016). Comprehensive genetic algorithm for ab initio global optimization of clusters. *Mol. Simul.* 42, 809–819. doi: 10.1080/08927022.2015.1121386

Zhao, S., Jin, R., Abroshan, H., Zeng, C., Zhang, H., House, S. D., et al. (2017). Gold nanoclusters promote electrocatalytic water oxidation at the nanocluster/cose2 interface. *J. Am. Chem. Soc.* 139, 1077–1080. doi: 10.1021/jacs.6b12529

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Chemistry

REVIEW
published: 26 November 2019
doi: 10.3389/fchem.2019.00809

Deep Learning for Deep Chemistry: Optimizing the Prediction of Chemical Patterns

Tânia F. G. G. Cova and Alberto A. C. C. Pais**

Coimbra Chemistry Centre, CQC, Department of Chemistry, Faculty of Sciences and Technology, University of Coimbra, Coimbra, Portugal

Computational Chemistry is currently a synergistic assembly between *ab initio* calculations, simulation, machine learning (ML) and optimization strategies for describing, solving and predicting chemical data and related phenomena. These include accelerated literature searches, analysis and prediction of physical and quantum chemical properties, transition states, chemical structures, chemical reactions, and also new catalysts and drug candidates. The generalization of scalability to larger chemical problems, rather than specialization, is now the main principle for transforming chemical tasks in multiple fronts, for which systematic and cost-effective solutions have benefited from ML approaches, including those based on deep learning (e.g. quantum chemistry, molecular screening, synthetic route design, catalysis, drug discovery). The latter class of ML algorithms is capable of combining raw input into layers of intermediate features, enabling bench-to-bytes designs with the potential to transform several chemical domains. In this review, the most exciting developments concerning the use of ML in a range of different chemical scenarios are described. A range of different chemical problems and respective rationalization, that have hitherto been inaccessible due to the lack of suitable analysis tools, is thus detailed, evidencing the breadth of potential applications of these emerging multidimensional approaches. Focus is given to the models, algorithms and methods proposed to facilitate research on compound design and synthesis, materials design, prediction of binding, molecular activity, and soft matter behavior. The information produced by pairing Chemistry and ML, through data-driven analyses, neural network predictions and monitoring of chemical systems, allows (i) prompting the ability to understand the complexity of chemical data, (ii) streamlining and designing experiments, (ii) discovering new molecular targets and materials, and also (iv) planning or rethinking forthcoming chemical challenges. In fact, optimization engulfs all these tasks directly.

Keywords: machine-learning, deep-learning, optimization, models, molecular simulation, chemistry

OPEN ACCESS

Edited by:
Emilio Martinez-Nunez,
University of Santiago de Compostela, Spain

Reviewed by:
Feng Wang,
Swinburne University of Technology, Australia
Anakuthil Anoop,
Indian Institute of Technology Kharagpur, India
Ruby Srivastava,
Centre for Cellular and Molecular Biology (CCMB), India

****Correspondence:***
Tânia F. G. G. Cova
tfirmino@qui.uc.pt
Alberto A. C. C. Pais
pais@qui.uc.pt

Specialty section:
This article was submitted to Theoretical and Computational Chemistry, a section of the journal Frontiers in Chemistry

Received: *16 July 2019*
Accepted: *11 November 2019*
Published: *26 November 2019*

Citation:
Cova TFGG and Pais AACC (2019) Deep Learning for Deep Chemistry: Optimizing the Prediction of Chemical Patterns. Front. Chem. 7:809. doi: 10.3389/fchem.2019.00809

INTRODUCTION

Patterns are ubiquitous in Chemistry. From the crystalline structures of solid forms to the branched chains of lipids, or the complex combinations of functional groups, chemical patterns determine the underlying properties of molecules and materials, essential to address important issues of societal concern. Artificial Intelligence (AI), and machine learning (ML) in particular, are committed to recognizing and learn from these patterns (Mitchell, 2014; Rupp, 2015; Goh et al., 2017; Li et al., 2017; Butler et al., 2018; Fleming, 2018; Gao et al., 2018; Kishimoto et al., 2018; Popova et al., 2018; Aspuru-Guzik et al., 2019; Elton et al., 2019; Gromski et al., 2019; Mater and Coote, 2019; Schleder et al., 2019; Venkatasubramanian, 2019).

Recent evidence on the most interesting and challenging prospects for accelerating discoveries in various chemistry fields, reported under "Charting a course for chemistry" (Aspuru-Guzik et al., 2019), indicate that the terms often used by the scientific community for describing the future trends in their field of research include "big data," "machine learning," and "artificial intelligence."

It is recognized that ML is already boosting computational chemistry, at different levels. Different aspects have been affected, and it is not easy to summarize developments in a consistent way. In what follows, the main areas in which ML has been employed are enumerated. These are extracted from recent contributions, that can be regarded as complementary and providing an overall perspective of the applications. These include different approaches for (i) understanding and controlling chemical systems and related behavior (Chakravarti, 2018; Fuchs et al., 2018; Janet et al., 2018; Elton et al., 2019; Mezei and Von Lilienfeld, 2019; Sanchez-Lengeling et al., 2019; Venkatasubramanian, 2019; Xu et al., 2019; Zhang et al., 2019), (ii) calculating, optimizing, or predicting structure-property relationships (Varnek and Baskin, 2012; Ramakrishnan et al., 2014; Goh et al., 2017; Simões et al., 2018; Chandrasekaran et al., 2019), density functional theory (DFT) functionals, and interatomic potentials (Snyder et al., 2012; Ramakrishnan et al., 2015; Faber et al., 2017; Hegde and Bowen, 2017; Smith et al., 2017; Pronobis et al., 2018; Mezei and Von Lilienfeld, 2019; Schleder et al., 2019), (iii) driving generative models for inverse design (i.e., produce stable molecules from a set of desired properties) (White and Wilson, 2010; Benjamin et al., 2017; Kadurin et al., 2017; Harel and Radinsky, 2018; Jørgensen et al., 2018b; Kang and Cho, 2018; Li et al., 2018b; Sanchez-Lengeling and Aspuru-Guzik, 2018; Schneider, 2018; Arús-Pous et al., 2019; Freeze et al., 2019; Jensen, 2019), (iv) screening, synthesizing, and characterizing new compounds and materials (Ahneman et al., 2018; Coley et al., 2018a; Granda et al., 2018; Segler et al., 2018; Li and Eastgate, 2019), (v) improving catalytic technologies and analytical tools (Li et al., 2017; Gao et al., 2018; Huang et al., 2018; Durand and Fey, 2019; Freeze et al., 2019; Schleder et al., 2019), (vi) developing quantum algorithms for molecular simulations, and (vii) progressing quantum sensing (Ramakrishnan et al., 2014; Ramakrishnan and Von Lilienfeld, 2017; Xia and Kais, 2018; Ahn et al., 2019; Christensen et al., 2019; Mezei and Von Lilienfeld, 2019; Zaspel et al., 2019; Zhang et al., 2019), just to name a few examples. In fact, Chemistry is a data-rich area, encompassing complex information which is often unstructured and poorly understood.

Deep learning (DL) approaches can also be particularly useful to solving a variety of chemical problems, including compound identification and classification, and description of soft matter behavior (Huang et al., 2018; Jha et al., 2018; Jørgensen et al., 2018b; Popova et al., 2018; Segler et al., 2018; Zhou et al., 2018; Chandrasekaran et al., 2019; Degiacomi, 2019; Elton et al., 2019; Ghosh et al., 2019; Mater and Coote, 2019; Matsuzaka and Uesawa, 2019; Xu et al., 2019).

The design of generalized cause/effect models, and the scaling-up of the contributions that are being made, containing high-dimensional data, and following the open-science basis (i.e., completely accessible, with precise metadata and practical formats) are critical challenges, that may, however, facilitate the routine implementation of data mining in chemistry and expedite new discoveries.

The amount and quality of chemical data generated by experiments and simulations have been the mainstay of the new data-driven paradigm, that establishes the bridge between theory, experiment, computation, and simulation.

This review describes, in a critical and comprehensive way, relevant contributions carried out recently and involving the development of chemistry ML approaches. An exhaustive account of the theoretical foundations and applications published in the early years of AI and ML in Chemistry falls beyond the scope of this review. The reader is referred to Lecun et al. (2015), Coveney Peter et al. (2016), Goh et al. (2017), Elton et al. (2019), Gromski et al. (2019), and Mater and Coote (2019) for a full description of these efforts.

Until 10 years ago, only a few 100 studies on the use of ML in Chemistry were published, resulting from the contributions made over four decades. In 2018, ca. 8,000 articles in the Web of Science database included these keywords, corresponding to an increase in ca. 35% for just one decade. In this review, there is room to mention only a small fraction of these applications.

Despite the increasing number of works on the topic, the models proposed and practices carried out by chemists are entailing serious concerns (Chuang and Keiser, 2018a). Several technical challenges, pitfalls, and potentials of ML, and also the reliability of the results, have been discussed by some authors (Ahneman et al., 2018; Chuang and Keiser, 2018a,b; Estrada et al., 2018) corroborating some critical remarks on the fragility of purely data-based approaches (Microsoft, 2018). "If data can speak for themselves, they can also lie for themselves." This reflects the need for an in-depth understanding of chemical patterns, data-driven and theory-driven models, and algorithms, before their application.

Although significant progress has been made connecting specific neural network predictions to chemical input features, understanding how scientists should analyze and interpret these models to produce valid and conclusive assumptions about the system under study, still remains to be fully defined.

Co-occurring Machine-Learning Contributions in Chemical Sciences

Scientific production covering ML-based approaches for dealing with chemical patterns has increased exponentially in recent years. However, the establishment and understanding of holistic, or macro insights on the major research trends in Chemistry sub-fields, are critical tasks. The challenge relies on how the analysis of these sub-fields, with thousands published works, reveals the most prominent applications supported by ML approaches (Butler et al., 2018; Chmiela et al., 2018; Chuang and Keiser, 2018a; Coley et al., 2018a; Gao et al., 2018; Lo et al., 2018; Panteleev et al., 2018; Xia and Kais, 2018; Ceriotti, 2019; Chan et al., 2019; Christensen et al., 2019; Gallidabino et al., 2019; Häse et al., 2019; Iype and Urolagin, 2019; Mezei and Von Lilienfeld, 2019; Schleder et al., 2019; Stein et al., 2019a; Wang et al., 2019).

In **Figure 1** an overview of the information generated during the last decade and ranked in the research domain of "Science Technology" of the Web of Science database, is presented.

The purpose of assessing the different facets of ML in Chemistry across the respective sub-fields is 3-fold: (i) to be able to quickly identify areas that have benefited most from the development and implementation of ML approaches, and those that still lack of such an optimization, as evidenced by the type of outcome, (ii) to identify the most relevant ML outcomes in each sub-field, and (iii) to assess the dynamics of ML outcomes over the 2008–2019 period and how these are related, giving rise to relevant research trends.

An extensive literature search on ML contributions in 30 Chemistry sub-fields is carried out, using a global set of 270 co-occurring keywords, each composed of three main terms, *machine learning, type of outcome* and the *sub-field* in which they co-occur (e.g., first co-occurrence: *Machine learning* AND *Quantum chemistry* AND *Quantum models*, second co-occurrence: *Machine learning* AND *Medicinal Chemistry* AND *Molecular screening*). A total of 5,279 contributions (including books, articles, reviews, editorials and letters) on ML in Chemistry, with 81,248 citations, and published between 2008 and June 30, 2019, are found in the worldwide Web of Science database, corresponding to a 4-fold increase over the previous four decades. Considering the compiled data and the selected Chemistry fields (organic, inorganic, physical, analytical, and biochemical), nine different ML outcomes embracing the most frequent chemical challenges are defined, including (i) text mining and system description, (ii) quantitative structure-activity/property relationships, (iii) DFT functionals and interatomic potentials, (iv) generative models and inverse molecular design, (v) molecular screening, (vi) synthesis/characterization of new compounds and materials, (vii) catalytic technologies, (viii) analytical techniques, and (ix) quantum models, algorithms, and quantum sensing. Note how these have a strong relation with the seven overall applications presented above (i–vii).

The heatmap represented in **Figure 1** reflects the impact of each type of ML outcomes on Chemistry sub-fields. The analysis of co-occurring keywords is thus performed in order to find the number of publications that appeared simultaneously in the selected sub-field. This relation is established with greater or lesser impact depending on the frequency of each set of keywords in the selected time-span.

The natural clusters generated from the most important co-occurring relationships are also identified. Considering the dendrogram for the Chemistry sub-fields, it can be observed that these are organized in two main groups, which discriminates, in general, classical Chemistry sub-fields (organic, inorganic, and physical) from analytical and biochemical sub-fields. This structure suggests a significant similarity in the type of ML outcomes within each group. Group 1 have benefitted from a significant production on catalytic technologies, DFT functionals and interatomic potentials, quantum models and quantum sensing. The most representative ML outcomes in group 2 are associated to text mining, analytical techniques, generative models and inverse design, molecular screening, structure activity relationships, and synthesis of new compounds and

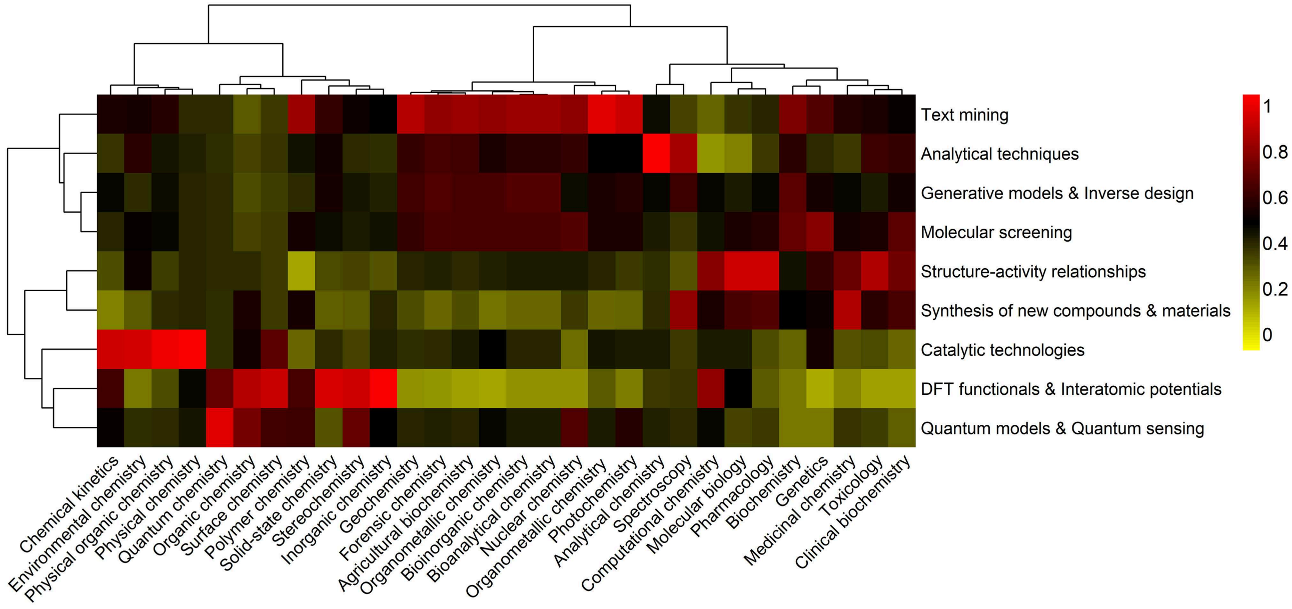


FIGURE 1 | A holistic view of ML-based contributions in Chemistry. The clustering heatmap displays the relative counts of ML outcomes, within each area of Chemistry (organic, inorganic, analytical, physical, and biochemistry), in the 2008–2019 (30 June) period. Data are expressed as fractions of the highest number of publications, including articles, reviews and books, containing specific co-occurring keywords, and following a standard normalization procedure. Hierarchical clustering with Euclidean distances and Ward linkage was performed on both Chemistry sub-fields and type of application. Co-occurrences are colored using a yellow-to-red color scheme. Highest and lowest relative contributions correspond to 1 (red) and 0 (yellow) values, respectively.

materials. Examination of the similarity between the type of ML outcomes reveals that there are three main groups, corresponding to (i) text mining, analytical techniques, generative modes and inverse design, and molecular screening (group 1), (ii) structure-activity relationships and synthesis of new compounds and materials (group 2), and (iii) catalytic technologies, DFT functionals and interatomic potentials, and quantum models and quantum sensing (group 3).

Historically, researchers have introduced numerical approximations to Schrödinger's equation, and the popular DFT calculations in *ab initio* approaches. However, the computational cost inherent to these classical approximations have limited the size, flexibility, and extensibility of the studies. Larger searches on relevant chemical patterns, have been successfully conducted since several research groups have developed ML models and algorithms to predict chemical properties using training data generated by DFT, which have also contributed to the increase of public collections of molecules coupled with vibrational, thermodynamic and DFT computed electronic properties (e.g., Behler and Parrinello, 2007; Rupp et al., 2012; Behler, 2016; Hegde and Bowen, 2017; Pronobis et al., 2018; Chandrasekaran et al., 2019; Iype and Urolagin, 2019; Marques et al., 2019; Schleder et al., 2019).

Based on the heatmap it can be determined that groups of Chemistry sub-fields have similar, but distinct ML-based contributions.

The increase in chemical data and scientific documents has boosted data mining and text mining processes to manage the huge amount of chemical information and to extract useful and non-trivial knowledge in different scenarios (Krallinger et al., 2017).

It is interesting to inspect if certain ML outcomes are produced in combination with each other.

In this context, the strongest correlation (0.97), shown in **Figure 2**, is observed between text mining and molecular screening, which is to be expected as a large number of molecules has been collected and screened systematically, by combining different text mining processes and chemoinformatics techniques (e.g., pharmacophore-based similarity and docking). These integrated approaches have allowed (i) extracting and collecting, in a systematic and high-throughput way, the available chemical and biological information from different sources (e.g., scientific documents) (Krallinger et al., 2017; Grzybowski et al., 2018), (ii) predicting activity based on chemical structure (Granda et al., 2018; Simões et al., 2018; Arús-Pous et al., 2019; Gromski et al., 2019; Lee et al., 2019; Li and Eastgate, 2019), and (iii) selecting promising molecular targets and candidates for further experimental validation (e.g., *in vitro* tests) (Ramakrishnan et al., 2014; Gupta et al., 2018; Segler et al., 2018; Brown et al., 2019; Elton et al., 2019; Li and Eastgate, 2019; Schleder et al., 2019; Xu et al., 2019).

Other strong correlations are found between generrative models & inverse design and the two abovementioned ML applications, molecular screening (0.95) and text mining (0.93). This can be explained by the fact that many researchers have proposed machine learning frameworks based on a variety of

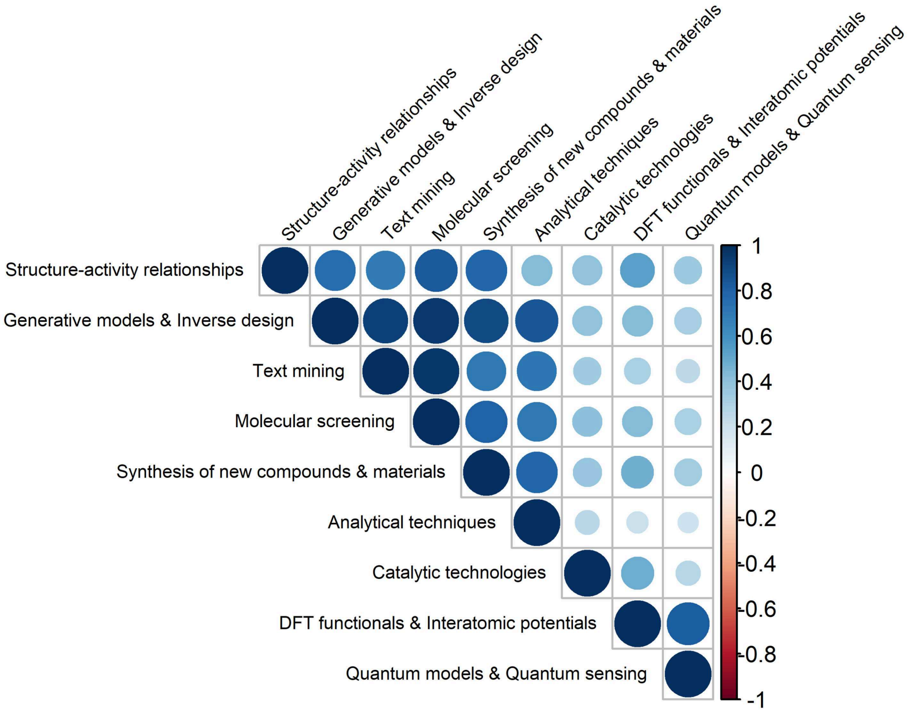


FIGURE 2 | Pairwise Pearson correlations between the different types of ML outcomes in Chemistry, produced in the 2008–2019 (30 June) period (darker colors reflect stronger correlations).

generative models for modeling molecules, which differ in the respective model structure and in the selected input features (Kadurin et al., 2017; Gupta et al., 2018; Jørgensen et al., 2018b; Arús-Pous et al., 2019; Brown et al., 2019; Jensen, 2019; Xu et al., 2019).

Also relevant are the correlations between generative models and inverse design and synthesis of new compounds and materials (0.90), and between generative models and inverse design and analytical techniques (0.85). The former relation evidences the significant effort that has been made on applying ML models, in particular those based on accurate DL architectures, to find and select lead molecules (e.g., drugs), displaying desired properties (Varnek and Baskin, 2012; Mitchell, 2014; Rupp, 2015; Lo et al., 2018). These properties are to be translated into a more simplified information on the molecular structures, and encoded into the respective chemical fingerprint (i.e., a set of binary characteristics of molecules). The process continues with the screening of the available databases for finding molecules that possess similar fingerprints to the generated ones. Generative models and deep neural networks (DLNs) have thus allowed generating molecules and promising candidates for useful drugs, basically from scratch, making it possible to "design perfect needles instead of searching for a needle in a haystack" (White and Wilson, 2010; Benjamin et al., 2017; Gómez-Bombarelli et al., 2018; Harel and Radinsky, 2018; Kang and Cho, 2018; Li et al., 2018b; Merk et al., 2018; Nouira et al., 2018; Popova et al., 2018; Sanchez-Lengeling and Aspuru-Guzik, 2018; Schneider, 2018).

It is also observed that there are other ML contributions that are interrelated: structure activity relationships with (i) molecular screening and (0.84), (ii) synthesis/characterization of new compounds and materials (0.78), and (iii) generative models and inverse design (0.75), DFT functionals and interatomic potentials with quantum models and quantum sensing (0.83), and synthesis/characterization of new compounds and materials with analytical techniques (0.79).

Both generative models and analytical techniques have been extensively used in the qualitative/quantitative search of patterns underlying chemical systems (Elton et al., 2019; Ghosh et al., 2019; Stein et al., 2019a,b). It should be noted the use data from large repositories (e.g., Protein Data Bank and Cambridge Structural Database) and ML methods are not new (Hiller et al., 1973; Gasteiger and Zupan, 1993; Behler, 2016). The latter have been employed as classification tools in pioneering works, encompassing, for e.g., the analysis of spectra (Thomsen and Meyer, 1989), quantification of structure-activity relationships (QSARs) (Agrafiotis et al., 2002), and prediction of binding sites of biomolecules (Keil et al., 2004).

The range of ML applications is now quite extended as a result of a deep integration of ML in analytical, theoretical and computational chemistry. Despite of some initial skepticism in understanding the foundations and structure of ML methods, their use has been accelerated and maturated in recent years essentially due to their suitability to new applications and industry needs, including chemical and pharmaceutical sectors.

MACHINE LEARNING FOR OPTIMIZATION: CHALLENGES AND OPPORTUNITIES

Designing models from chemical observations to study, control, and improve chemical processes and properties is the basis of optimization approaches. The understanding of chemical systems, and the respective underlying behavior, mechanisms and dynamics, is currently facilitated by the development of descriptive, interpretative, and predictive models, i.e., approximations that represent the target system or process. Applications of such models have included the (i) optimization of reaction parameters and process conditions, e.g., changing the type of reagents, catalysts, and solvents, and also varying systematically, concentration, addition rate, time, temperature, or solvent polarity, (ii) suggestion of new reactions based on critical functional groups, (iii) prediction of reaction/catalyst design, and optimization of heterogeneous/homogeneous catalytic reactions, (iv) acceleration and discovery of new process strategies for batch reactions, (v) establishment of trade-offs in the reaction rate and yield of organic compounds, (vi) description and maximization of the production rate and conversion efficiency of chemical reactions, (vii) prediction of the potential toxicity of different compounds, and also the (viii) rational design of target molecules and guided exploration of chemical space (Kowalik et al., 2012; Houben and Lapkin, 2015; Houben et al., 2015; Zielinski et al., 2017; Häse et al., 2018; Min et al., 2018; Zhou et al., 2018; Ahn et al., 2019; Choi et al., 2019; Gromski et al., 2019; Matsuzaka and Uesawa, 2019).

ML provides the tools to scrutinize and extract useful information to be employed in modeling and system-solving solutions (Artrith and Urban, 2016; Ward and Wolverton, 2017). In Chemistry domains, researchers have had access to multidimensional data of unprecedented scale and accuracy, that characterize the systems/processes to be modeled. A collection of different examples of optimization based on ML approaches can be found in Kowalik et al. (2012), Houben and Lapkin (2015), Houben et al. (2015), Cortés-Borda et al. (2016), Wei et al. (2016), Benjamin et al. (2017), Ahneman et al. (2018), Gao et al. (2018), Granda et al. (2018), Min et al. (2018), Ahn et al. (2019), Elton et al. (2019), Matsuzaka and Uesawa (2019).

Specifically, ML contributions have involved a variety of systems including drugs (Griffen et al., 2018), polymers (Li et al., 2018a), polypeptides (Grisoni et al., 2018; Müller et al., 2018), energetic materials (Elton et al., 2018), metal organic frameworks (He et al., 2018; Jørgensen et al., 2018a; Shen et al., 2018), and organic solar cells (Jørgensen et al., 2018a).

Advances in analytical methods, laboratory equipment and automation have rapidly improved the performance of experimental procedures (e.g., miniaturizing experiments for reactions, and connecting analytical instruments to advanced software based on decision-making algorithms and optimization tools) (Stevens et al., 2010; Smith et al., 2011; Richmond et al., 2012; Houben and Lapkin, 2015). The implementation of ML-based approaches have allowed developing innovative capabilities, such as cost-effective experiments, advanced algorithms for automation, and designing of experiments, chemoinformatics tools for dealing with high-dimensional analytical data, and accelerated *in situ*/in line analysis of chemical transformations (e.g., polymerization reactions, heterogeneous catalytic processes, aggregation of nanoparticles) (Houben and Lapkin, 2015; Häse et al., 2018).

However, there are critical challenges that ML in Chemistry must face, including the control of experiments, the detailed description of chemical space, the flexibility and generalization of models, robustness, and accuracy of predictions, and the establishment of effective cross-disciplinary collaborations (Montavon et al., 2013; Hansen et al., 2015; Kishimoto et al., 2018; Smith et al., 2018a).

A clear definition of ML, as well as the distinction from other purely mathematical regression methods is not straightforward, and can be associated to some degree of arbitrariness (Behler, 2016). Standard ML methods include, artificial neural networks, support vector machines, and Gaussian processes, which have contributed to the rational design of compounds and materials, and to the improvement of computational frameworks (Goh et al., 2017; Mater and Coote, 2019). The latter have been applied for e.g., in QSAR models and drug design (Kadurin et al., 2017; Chen et al., 2018; Fleming, 2018; Green et al., 2018; Gupta et al., 2018; Li et al., 2018b; Lo et al., 2018; Popova et al., 2018; Simões et al., 2018) aiming at identifying systems, molecules and materials with optimal properties (e.g., conductivity, aqueous solubility, bioavailability, bioactivity, or toxicity) (Kadurin et al., 2017; Freeze et al., 2019). This can be made via extensive searches, in large databases, of latent relationships between the atomic structures. The structures, can thus be encoded using multiple descriptors, and target properties.

The possibilities of applying ML for optimization in Chemistry are endless. There are studies focused on ML approaches for inferring on the optimized geometry of a system (Zielinski et al., 2017; Venkatasubramanian, 2019), and finding minima on complex potential energy surfaces (Chen et al., 2015; Chmiela et al., 2018; Kanamori et al., 2018; Xia and Kais, 2018; Hughes et al., 2019), such as those of large water clusters (Bose et al., 2018; Chan et al., 2019).

The most innovative aspects of ML in Chemistry are related to the availability of large volumes of theoretical data (e.g., electrostatic energy contributions in force fields, atomic charges, structural properties, and representations of the potential energies), obtained from automatic and accurate electronic structure calculations (Behler, 2016).

However, the intricate nature of the configuration space and its exponential dependence on system size and composition, have hampered the screening of the entire set of candidate structures directly by electronic structure calculations (Behler, 2016; Welborn et al., 2018).

Signs of Controversy

Despite the usefulness of ML approaches being indisputable, with the promise to modernize molecular simulations, synthesis, materials science, and drug discovery, the respective endorsement and practical aspects in some chemical sub-fields is far from consensual (Ahneman et al., 2018; Chuang and Keiser, 2018a,b).

Ten years ago, there were only a few publications on applications of ML in Chemistry, but currently there are thousands of published works. The controversy has highlighted the potential (instructive) pitfalls of some practices using ML. It has been argued that ML algorithms may lead to overestimated performances and deficient model generalizations, due to their sensitivity to the presence of maze-like variables and experimental artifacts (Chuang and Keiser, 2018a). For instance, Ahneman et al. (2018) have recently designed a ML model to predict yields of cross coupling reactions with high accuracy, containing isoxazoles, as reaction inhibitors, which were incorporated for assessing the robustness of the reaction. Input data for the proposed algorithm included yields and reagent parameters of 3,000 reactions, such as NMR shifts, dipole moments, and orbital energies. The most significant features of the proposed algorithm were found to be the descriptors of additives. However, the experimental design of this original work has been contested by Chuang and Keiser (2018b), who warned for potential artifacts associated to the original work. These authors demonstrated that the model also identified reaction additives as the descriptors displaying the greatest impact on the reactions, suggesting that high additive feature contributions cannot be discriminated from the hidden structure within the dataset, i.e., the procedure of the original paper was not sufficient for establishing isoxazole additives as the most important descriptors (Chuang and Keiser, 2018b). A meticulous preprocessing of input data and validation of the model hypothesis was then suggested. The Y-randomization test in the original work was taken into account just the information rooted in the structure of the data set, irrespective to the intended outcome. The classical approach based on multiple hypotheses to assess alternative descriptions of the performance of the ML model was implemented (Chuang and Keiser, 2018b). The effect of different reaction parameters (e.g., additives, catalyst, and aryl halide) in an extensive combinatorial layout generated over several independent reactions was duly explored, providing the underlying structure of the data (Chuang and Keiser, 2018b).

An alternative assumption considering that ML algorithms deal with patterns within the experimental design, instead of learning from the most relevant chemical features was therefore investigated. It was concluded that ML is prone to explore data irrespective to their size and structure. This aspect was illustrated by extracting and replacing the chemical features (e.g., electrostatics, NMR shifts, dipole moments) from each molecule with random (Gaussian distributed) numeric strings. It was shown that the predictions were similar to the original ones. Chuang and Keiser (2018a) have also introduced technical and conceptual standpoints, including the use of adversarial controls to evaluate the predictive performance of ML models, focusing on the design of rigorous and deliberated experiments, ensuring accurate predictions from suitable and significant models (Chuang and Keiser, 2018a). By revising the original information, a number of variations of the test sets was introduced by Estrada et al. (2018) for assessing the performance of predictions, considering alternatives to the random-forest model. It was therefore demonstrated that ML models are in fact quite sensitive to such imposed features, and the reagent-label models are relevant representations of the data set and useful for comparing performances in generalization assessments.

The original assumptions regarding the significance and validity of the random-forest (chemical-feature) model to describe important and general chemical features were also confirmed (Estrada et al., 2018).

A lesson that chemists may draw from such constructive discussions is that as the size of the data set increases, the performance of ML models also increases, but with the possibility of obtaining unexpected results and irrelevant patterns, as the rules for ML algorithms to detect and deal with potential technical and conceptual gaps are not well-established. Specifically, the description of chemical reactivity underlying a data set is required in order to ensure the reaction prediction, by using data and reagent-label models to evaluate the scope and restraints of chemical characterization.

ML provides new opportunities to increase the quality and quantity of chemical data, which are essential to promote optimization, implementation of rational design and synthetic approaches, prioritization of candidate molecules, decision-making, and also for guiding of innovative ideas.

Deep Learning, Deep Chemistry

In this section, an introductory overview into the core concepts of DL, and DLNs is provided. Focus is given to the unique properties of DL, that distinguish these algorithms from traditional machine learning approaches, with emphasis on chemical applications rather than providing theoretical and mathematical details.

ML is a branch of computer science dedicated to the development of algorithms capable of learning and making decisions on complex data (Samuel, 1959; Mitchell, 1997). This learning process involves specific tasks that are commonly classified in (i) supervised learning, for establishing the relationship between input and output data (e.g., linear regressions and classification techniques), (ii) unsupervised learning, for finding hidden patterns or features in data, without any previous information on such characteristics and interrelations (e.g., clustering and dimension reduction techniques), and (iii) reinforcement learning, for performing a particular task through repeated dynamic interactions e.g., optimization of molecules (Zhou et al., 2018) and chemical reactions (Zhou et al., 2017).

Deep learning is a fast-moving sub-area of ML, focused on sophisticated learning and extrapolation tasks, fostered by the wide range of chemistry literature, open-source code, and datasets (Goh et al., 2017).

The ability of DL to establish the relevant phenomena, expedite chemical reactions, and predict relevant properties, optimal synthesis routes, solve critical analytical uncertainties, and reduce costs and resources, is invaluable in Chemistry. Its success in modeling compound properties and reactions, depends, among other aspects, on the access to

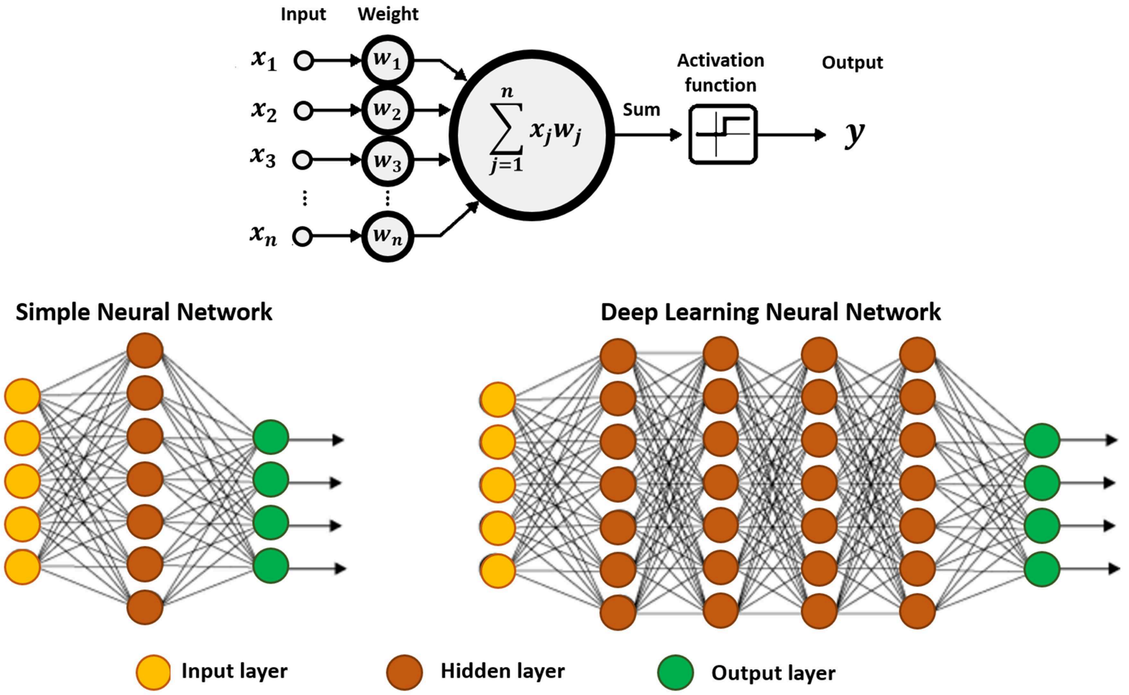


FIGURE 3 | Schematic representation of an artificial neuron (top), and a simple neural network displaying three basic elements: input, hidden and output layers (bottom-left), and a deep neural network showing at least two hidden layers, or nodes (bottom-right). The calculation is performed through the connections, which contain the input data, the pre-assigned weights, and the paths defined by the activation function. If the result is far from expected, the weights of the connections are recalibrated, and the analysis continues, until the outcome is as accurate as possible.

comprehensive, historical repositories of published chemical data (Venkatasubramanian, 2019).

There are barriers to be surpassed, including cleaning data, production of meaningful and accurate chemical information (free of bias), lack of standardization of chemical data, expertise and familiarity with ML and DL in chemistry sectors, and also lack of collaboration opportunities) (Mater and Coote, 2019).

The majority of DL algorithms currently developed are based on artificial neural networks (Lecun et al., 2015).

DLNs are now a proving-ground for research in chemical sciences (Goh et al., 2017; Jha et al., 2018; Popova et al., 2018; Segler et al., 2018; Elton et al., 2019; Mater and Coote, 2019; Xu et al., 2019). Similarly to artificial neural networks, DLNs are produced to resemble the brain, in which the information passes through a series of interconnected nodes comparable to neurons (Lecun et al., 2015). Each node analyzes segments of information and transfer that information to adjacent nodes (see **Figure 3**).

The computational model consists of multiple hidden layers (in higher number comparing to more conventional approaches) which confer the ability of DLNs to learn from highly complex data and perform correlation and reduction. This means that the algorithm discovers correlated data, while discarding irrelevant information. Each layer combines information collected from the previous layer, and subsequently infers on the respective significance and send the relevant information to the next layer. The hidden term is used to represent layers that are not direct neighbors of the input or output layers.

The process allows constructing increasingly complex and abstract features, by adding layers and/or increasing the number of neurons per layer. However, the use of more than a single hidden layer requires determining error attributions and corrections to the respective weights. This is carried out via a backpropagation, i.e., a backward process starting from the predicted output, and back through the neural network (Goh et al., 2017). In this process a gradient descent algorithm is employed to determine the minimum in the error surface created by each respective neuron, when generating the output. Note that, this gradient descent approach is conceptually similar to the steepest descent algorithm implemented in classical MD simulations (Goh et al., 2017). The major difference lies on the use of an iterative process, in which an error function of the target output of the neural network is minimized, and the weights of the neurons are updated, instead of iteratively minimizing an energy function and updating atomic coordinates for each step.

A complete description of the main core concepts and architecture of DL applied to chemistry is given in Goh et al. (2017) and Mater and Coote (2019).

Other interesting reviews covering theoretical aspects (Goh et al., 2017), available descriptors and datasets, and also comparing model performances (Wu et al., 2017) have been published. Moreover, a wide range of ML applications, including drug design (Ekins, 2016; Chen et al., 2018; Fleming, 2018), synthesis planning (Coley et al., 2018a), medicinal chemistry (Panteleev et al., 2018), cheminformatics (Lo et al., 2018), quantum mechanical calculations (Rupp, 2015), and materials science (Butler et al., 2018) have been collected.

A summary of the main contributions of DL for solving relevant chemical challenges, as well as an illustration of the general components of a DL framework are presented in **Figure 4**.

DL algorithms are particularly attractive for accelerating discoveries in pharmaceutical, medicinal and environmental chemistry (El-Atta and Hassanien, 2017; Goh et al., 2017; Klucznik et al., 2018; Miller et al., 2018; Panteleev et al., 2018; Smith et al., 2018b; Wu and Wang, 2018; Molga et al., 2019), since they have made possible, for e.g., to simulate millions of toxic compounds and identify those compounds displaying target properties, safely, economically, and sustainably. These types of applications have been thoroughly revised in various publications and will not be further addressed in what follows [see for e.g., (Kadurin et al., 2017; Chen et al., 2018; Fleming, 2018; Green et al., 2018; Gupta et al., 2018; Li et al., 2018b; Lo et al., 2018; Panteleev et al., 2018; Popova et al., 2018; Smith et al., 2018b)].

DL is not only a cost-cutting effort, but also an innovative source of new perspectives.

CUTTING-EDGE APPLICATIONS

In recent years, ML has been evoked in chemistry-related tasks. The use of ML and, in particular, DL-based approaches across prediction of binding, activity and other relevant molecular properties, compound/material design and synthesis, as well as applications of genetic algorithms are highlighted in what follows.

Researchers in chemical sciences have started exploring the capabilities of ML using data collected from computations and experimental measurements. Data mining is traditionally adopted to explore high-dimensional data sets, in order to identify and establish relevant connections for the chemical features of compounds and materials.

Other more ambitious approaches, including quantum mechanics, which integrates physics-based computations (e.g., DFT) and ML methods in the search for novel molecular components, have also been implemented (Curtarolo et al., 2013).

Amongst the major achievements of DL in Chemistry, are the outstanding performances in predicting activity and toxicity, in the context of the Merck activity prediction challenge in 2012, and the Tox21 toxicity prediction challenge launched by NIH in 2014, respectively. In the former, DL was very successful in the competition outperforming Merck's internal baseline model. In the second challenge, DL models also achieved top positions (Goh et al., 2017).

Similarly to what happens to the majority of the modern computational chemists who no longer build their own code to perform MD simulations or quantum chemical calculations, due to the existence and availability of well-established software packages, DL researchers have also use several software packages for training neural networks including Torch, Caffe, Theano, and Tensorflow (Goh et al., 2017).

Apart from the influence of software improvements, the continuous growth of chemical data in public databases, such as PubChem and Protein Data Bank has also facilitated the

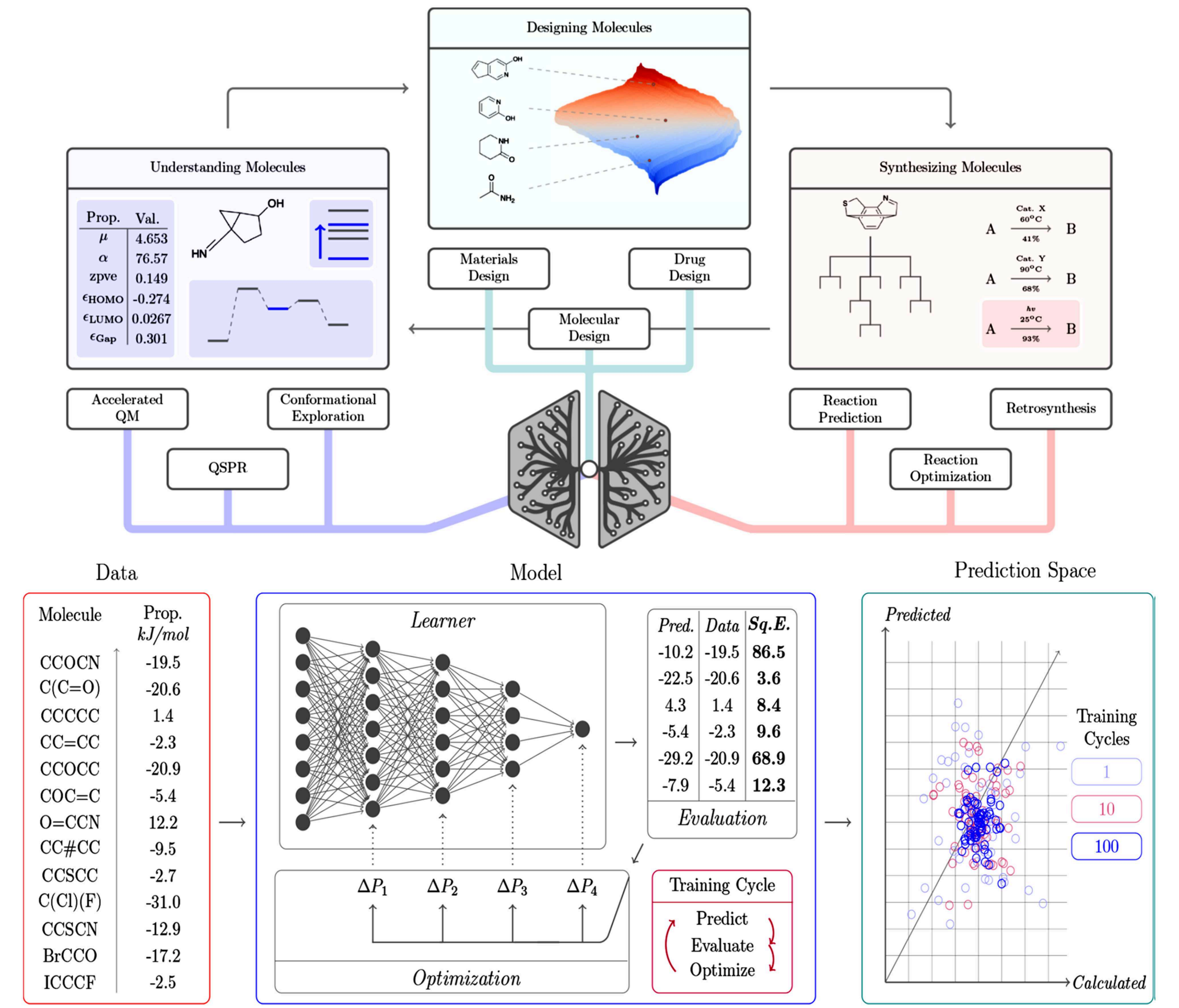


FIGURE 4 | Overview of (top) the contribution of DL algorithms for solving different chemical challenges and the respective tasks, and (bottom) the general components of a DL framework, including the input data, the learning model able to interpret the data and the prediction space, from which the model performance can be inspected. The model represents an optimization cycle containing interconnected components: prediction, evaluation, and optimization. Reprinted with permission from Mater and Coote (2019). Copyright (2019) American Chemical Society.

raise of ML and DL applications in Chemistry, including quantum chemistry, property prediction and materials design, drug discovery, QSAR, virtual screening, and protein structure prediction (Goh et al., 2017; Christensen et al., 2019).

Improving Computational and Quantum Chemistry

Computational chemistry is naturally a sub-field that has been increasingly boosted by the advances and unique capabilities of ML (Rupp et al., 2012; Ramakrishnan et al., 2014, 2015; Dral et al., 2015; Sánchez-Lengeling and Aspuru-Guzik, 2017; Christensen et al., 2019; Iype and Urolagin, 2019; Mezei and Von Lilienfeld, 2019; Zaspel et al., 2019).

Also, recent progresses have enabled the acceleration of MD simulations (atomistic and coarse-grained), contributing to increase knowledge on the interactions within quantum many-body systems and efficiency of DFT-based quantum mechanical modeling methods (Bartók et al., 2010, 2013; Behler, 2011a,b, 2016; Rupp et al., 2012, 2015; Snyder et al., 2012; Hansen et al., 2013, 2015; Montavon et al., 2013; Schütt et al., 2014; Alipanahi et al., 2015; Botu and Ramprasad, 2015b; De et al., 2016; Faber et al., 2016; Sadowski et al., 2016; Wei et al., 2016; Brockherde et al., 2017; Chmiela et al., 2017, 2018; Smith et al., 2017; Wu et al., 2017; Gómez-Bombarelli et al., 2018). This field is still in its infancy and have offered invaluable opportunities for dealing with a wide range of challenges and unsolved questions, including but not limited to model accuracy, interpretability, and causality.

For instance, the prediction of the refractive index of ionic liquids based on quantum chemistry calculations and an extreme learning machine (ELM) algorithm has been conducted (Kang et al., 2018). Specifically, 1,194 experimental data points for 115 ionic liquids at different temperatures were collected from more than 100 literature reports. Quantum chemistry calculations were performed for obtaining the structures and descriptors of the ionic liquids. The model was designed using a stepwise regression algorithm and the R^2 and AARD% values were 0.841 and 0.855%, respectively. It was found that prediction of the refractive index was significantly affected by ionic liquid anions, comparing to the cations. Better performances were achieve using the ELM algorithm, with the R^2 and AARD% values of 0.957 and 0.295%, respectively (Kang et al., 2018).

ML has also contributed for modeling the water behavior, shedding light on important phenomena related to water molecules interactions and the resulting density. Morawietz et al. (2016) have calculated ice's melting point from fundamental quantum mechanics, demonstrating the predictive power of ab initio MD simulations and highlighting the critical role of van der Waals forces (Morawietz et al., 2016). It was evidenced that ice occupies a larger volume than liquid water as hydrogen bonds display water molecules in a rigid 3D network. These hydrogen bonds weaken when ice melts, and water molecules approximate, becoming dense with an extreme value at 4°C (Morawietz et al., 2016). Note that these processes can also be rationalized resorting to *ab initio* MD approaches based on DFT; however, such calculations are associated to highly demanding computations. In addition to this, DFT approaches are not able to accurately reproduce minute but relevant van der Waals forces. The same authors have trained a neural network to reproduce DFT results with less computer power, and employed a previously-existing van der Waals correction. Water density changes, hydrogen bond network flexibility, and competition effects in terms of the nearest shell's contraction, after cooling, were explained based on the simulations (Morawietz et al., 2016).

One of the current challenges is to answer the question of whether chemical-physical properties, that often require quantum mechanics (e.g., dipole moments, binding and potential energies, and thermodynamics), can be represented and predicted by ML methods (Hansen et al., 2013, 2015; Montavon et al., 2013; Faber et al., 2016; Iype and Urolagin, 2019; Jaquis et al., 2019). Several attempts have been made on the topic with some successful examples (Rupp et al., 2012; Faber et al., 2017).

Rupp et al. (2012) have developed a model based on nuclear charges and atomic positions for predicting molecular atomization energies of various organic compounds. A matrix composed of molecular elements and configuration was built, describing the potential energy of each individual atom and the Coulomb repulsion between nuclear charges. A non-linear regression scheme was employed for solving and mapping the molecular Schrödinger equation.

The regression models were trained and compared to atomization energies calculated with hybrid DFT, transforming a 1-h run (on average) of hybrid DFT per each atomization energy into milliseconds using ML. Cross-validation over more than seven thousand organic molecules yielded a mean absolute error below 10 kcal/mol. The authors have trained the ML algorithm on a set of compounds in a database, comparing the respective matrices to determine differences between molecules, so as to develop a landscape of such differences. Based on the atomic composition and configuration, the unknown molecule can be positioned in the landscape and the respective atomization energy can be estimated by the contributions (weights) obtained from the differences between the unknown and all known molecules (Rupp et al., 2012).

More recently, the impact of selecting regressors and molecular representations on the construction of fast ML models of several electronic ground-state properties of organic molecules has also been investigated (Faber et al., 2017). The performance of each combination between regressor, representation, and property was evaluated with learning curves, which allowed reporting out-of-sample errors, as a function of the size if the training set (ca. 118 k molecules). The QM9 database (Ramakrishnan et al., 2014) was used for extracting the molecular structures and properties at the hybrid DFT level of theory, and included data on dipole moment, polarizability, enthalpies and free-energies of atomization, HOMO/LUMO energies and gap, heat capacity, zero point vibrational energy, and the highest fundamental vibrational frequency.

Several regression methods including linear models (Bayesian ridge regression and elastic net regularization), random-forest, kernel ridge regression, and neural networks (graph convolutions and gated graph networks) were tested. It was concluded that out-of-sample errors were strongly affected by the molecular properties, and by the type of representation and regression method. Molecular graphs and graph convolutions displayed better performances for electronic properties, while kernel ridge regression and histograms of dihedrals were suitable for describing energy-related properties [see Faber et al. (2017) for details on other relevant combinations]. Predictions based on the ML model for all properties have shown lower deviations from DFT (B3LYP) than the latter deviated from experiment. ML models displayed thus an improved prediction accuracy than hybrid DFT, since experimental or explicitly electron correlated quantum data were available.

In terms of drug development Brockherde et al. (2017) have developed a ML algorithm for predicting the behavior of molecules with potential to be used as pharmaceuticals and in the design of new molecules, able to enhance the performance of emerging energetic materials, including solar cells, battery technologies, and digital displays. The main goal was to identify the underlying patterns in the molecular behavior, by employing the ML algorithm for understanding atomic interactions within a molecule and using such information to predict new molecular phenomena.

Specifically, the algorithm was created and trained on the basis of a small sample set of the molecule under study, and applied to simulate the intricate chemical behavior within selected molecules, including malonaldehyde. A directed learning of the density-potential and energy-density maps was conducted, as illustrated in **Figure 5**, and the first MD simulation of with a ML density functional on malonaldehyde was performed, allowing to describe the intramolecular proton transfer process (Brockherde et al., 2017).

In more detail, one of the key tasks in atomistic modeling is the prompt and automated analysis of the simulation results, in order to provide a comprehensive understanding of the behavior of individual atoms and target collective properties. The main supervised and unsupervised machine-learning methods directed at classifying and coarse-graining of molecular simulations were recently summarized and discussed in Ceriotti (2019). A schematic overview of these methods, and also of a workflow reflecting the application of a ML scheme to an atomic-scale system is presented in **Figure 6**.

Also relevant is the development of improved molecular force fields, commonly used in MD simulations, using ML. On the other hand, the intrinsic operational aspects of MD simulations, in which the dynamic evolution of the chemical system is detailed in a fixed period of time, and for which interparticle forces and potential energies are often estimated using interatomic potentials, or molecular mechanics force fields, are perfectly suited for ML. In fact, some of the timesteps can be used as a training phase for estimating consecutive ones, assuming that each of the timesteps of MD simulation is strongly correlated with the preceding timestep and is adequate for sampling the phase space rapidly and accurately, allowing to estimate any meaningful property (Behler, 2016). MD simulations often sample abnormal, but probably relevant configurations, requiring the implementation of a decision tool for dealing with the unusual configuration, and from which ML may turn off and start learning (Botu and Ramprasad, 2015a; Smith et al., 2018a). These conditions have also been previously discussed and applied to *ab initio* MD (Botu and Ramprasad, 2015a).

In MD, the energies and forces for a vast number of atomic configurations are required, which can be obtained by performing the electronic structure calculations along the trajectory, or by evaluating the direct functional relation between the atomic configuration and the energy (Mansbach and Ferguson, 2015). This analytic expression, defined before running the simulation, is often recognized as a force field, an interatomic potential, or a potential-energy surface. Calculations of electronic structures are very demanding, even for DFT. DFT-based *ab initio* MD simulations are restricted to a few 100 atoms and shorter simulation times (Ahn et al., 2019).

The requirements for calculating ML potentials are very similar to conventional empirical potentials, and are duly discussed in Behler (2016). More recent conventional force fields are developed and validated for very specific systems, being limited by the functional form upon which they were constructed. On the other hand, despite requiring a training set, ML-based force fields are adaptive and more robust upon configurations not previously sampled (Botu and Ramprasad, 2015a). Furthermore, these force fields can be extended rapidly to different types of atoms and molecules, as they can learn and apply the physical laws, rather than starting from strarch (Botu et al., 2017).

Several improved force fields, and accurate predictions of thermodynamics and kinetics signatures, as well as their influence in molecular structures have been provided by performing ML-based atomistic and *ab initio* MD simulations. For instance, Chmiela et al. (2018) have incorporated spatial and temporal physical symmetries into a gradient-domain machine learning (sGDML) model for constructing flexible molecular force fields from high-level *ab initio* calculations, with a great potential to be used to improve spectroscopic accuracy in molecular simulations. The sGDML model was able to reproduce global force fields at quantum-chemical CCSD(T) level of accuracy and produced converged MD simulations with fully quantized electrons and nuclei (Chmiela et al., 2018).

The parameterization of force fields and semiempirical quantum mechanics have also been performed integrating ML and evolutionary algorithms (Wang et al., 2019), which were successfully applied in MD (Wang et al., 2019). Constructing coarse-grained molecular models has been a common approach to extend the time/length-scales accessible to large or complex systems (Wang et al., 2019). These models have allowed establishing suitable interaction potentials for properties of high-resolution models or experimental data. Wang et al. (2019) have reformulated coarse-graining as a supervised machine learning problem, by using statistical learning theory for decoupling the coarse-graining error, and cross-validation for choosing and comparing the performance of distinct models. For that purpose, the authors developed a DL model, that learned coarse-grained free-energy functions and was trained by a force-matching strategy (see **Figure 7**).

The proposed framework automatically learned multiple terms necessary for accurate coarse-grained force fields, i.e., was able to keep relevant invariances and incorporate physics knowledge, avoiding the sampling of unphysical structures.

The class of coarse-grained directed neural networks can thus be trained with the force-matching principle and can encode all physically relevant invariances and constraints, including invariance of (i) the free-energy and mean force with respect to translation of the molecule, (ii) the free-energy and variance of the mean force associated to molecular rotation, and considering (iii) the mean force being a conservative force field generated by the free-energy, and (iv) a prior energy for preventing deviations of the simulations with coarse-grained neural networks into unphysical state space regions, i.e., states displaying overstretched bonds or clashing atoms, which are captured out of the training data.

The proposed strategy also outperformed classical coarse-graining approaches, which generally failed to capture relevant features of the free-energy surface, providing the all-atom explicit-solvent free-energy surfaces estimated with models including just a few coarse-grained beads, in the absence of solvent (Wang et al., 2019).

The integration of ML in MD simulations have also been useful for understanding the rate and yield of chemical reactions and providing key mechanistic details (Christensen et al., 2019; Häse et al., 2019). For instance, an unsupervised ML analysis

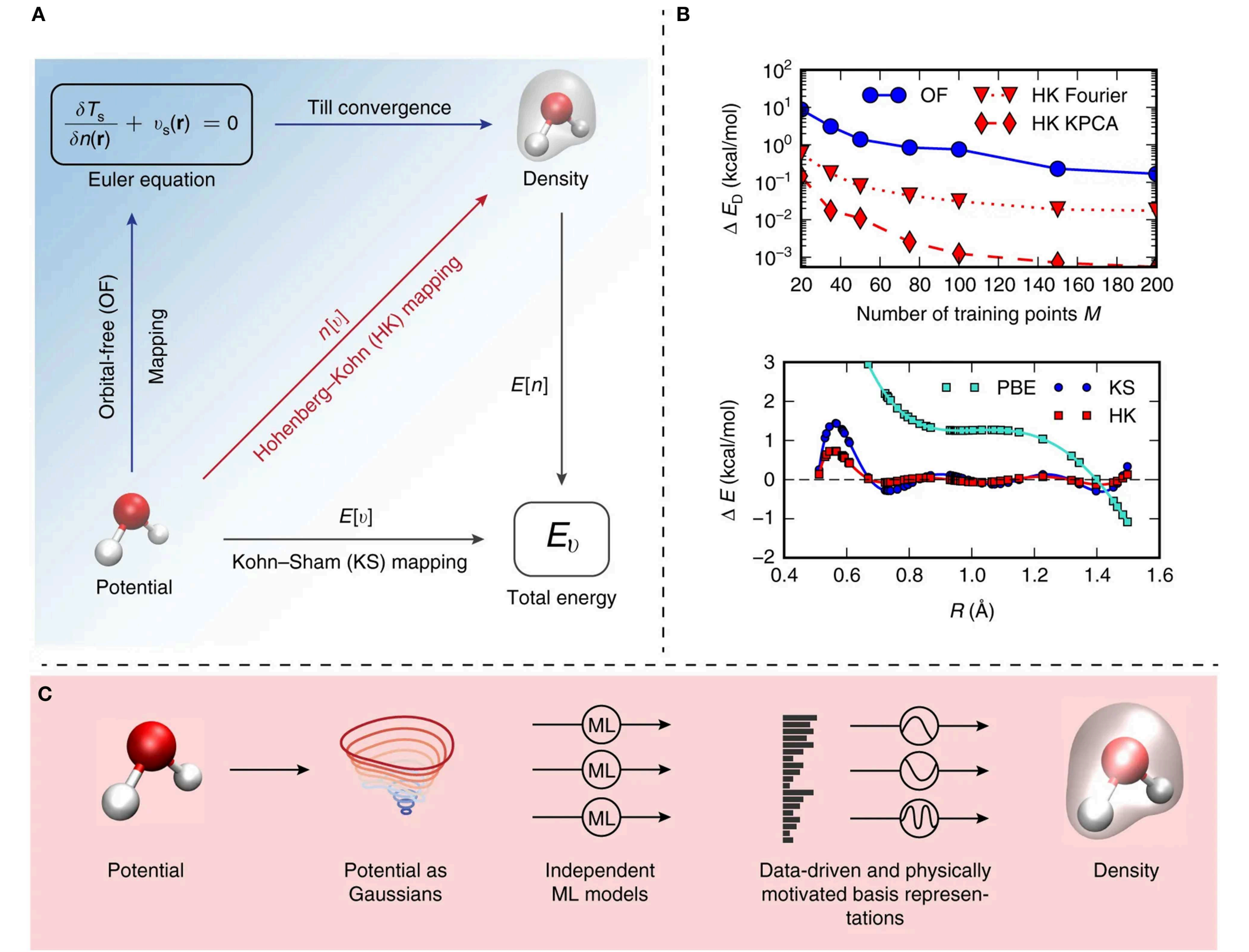


FIGURE 5 | (A) Illustrative summary of the mappings proposed by Brockherde et al. (2017). $E[v]$ is a conventional electronic structure calculation, i.e., Kohn–Sham density functional theory (KS-DFT) and is represented by the bottom vector. The ground-state energy is determined by solving KS equations given the external potential, v. $E[n]$ corresponds to the total energy density functional. The Hohenberg–Kohn map n[v] (red vector) from external potential to its ground state density is also presented. **(B)** Top: graphical representation of the dependency of the energy error on the number of training points (M), for ML-OF and ML-HK, considering different basis sets for the one-dimensional problem. Bottom: errors in the Perdew-Burke-Ernzerhof (PBE) energies and the ML maps as a function of interatomic spacing, R, for H_2 with M = 7. **(C)** Schematic illustration of the strategy for obtaining predictions based on the proposed machine learning Hohenberg–Kohn (ML-HK) map. Molecular geometry is represented by Gaussians, several independent Kernel ridge regression models allows predicting each basis coefficient of the density. The performance of data-driven (ML) and common physical basis representations for the electron density is assessed.

tool based on Bayesian neural networks (BNNs) was proposed by Häse et al. (2019) to extract relevant information from *ab initio* MD simulation of chemical reactions (Häse et al., 2019). BNNs have been optimized to predict a specific outcome of an *ab initio* MD simulation corresponding to the dissociation time of the unmethylated and tetramethylated 1,2-dioxetane molecules, from the initial nuclear geometry and velocities. Predictions based on BNNs showed that an earlier dissociation was related to the planarization of the two formaldehyde moieties and also to the symmetric shortening of the C–O bonds, respecting the octet rule, i.e., the relation between bond order and bond length and orbital hybridization (Häse et al., 2019).

Rupp et al. (2012) have developed a ML algorithm based on non-linear statistical regression to predict the atomization energies of organic molecules. The proposed model employed a subset of seven thousand elements of the database, and a library of more than 100 stable and synthetically-tractable organic compounds. The target data used to train the model included atomization energies of the compounds calculated using the PBE0 hybrid functional. Cartesian coordinated and nuclear charge were used as descriptors in a "Coulomb" matrix representation. A mean-absolute error accuracy of 14.9 kcal/mol was achieved using a small fraction of the compounds for the training set. Similar accuracy, ca. 15.3 kcal/mol, was obtained

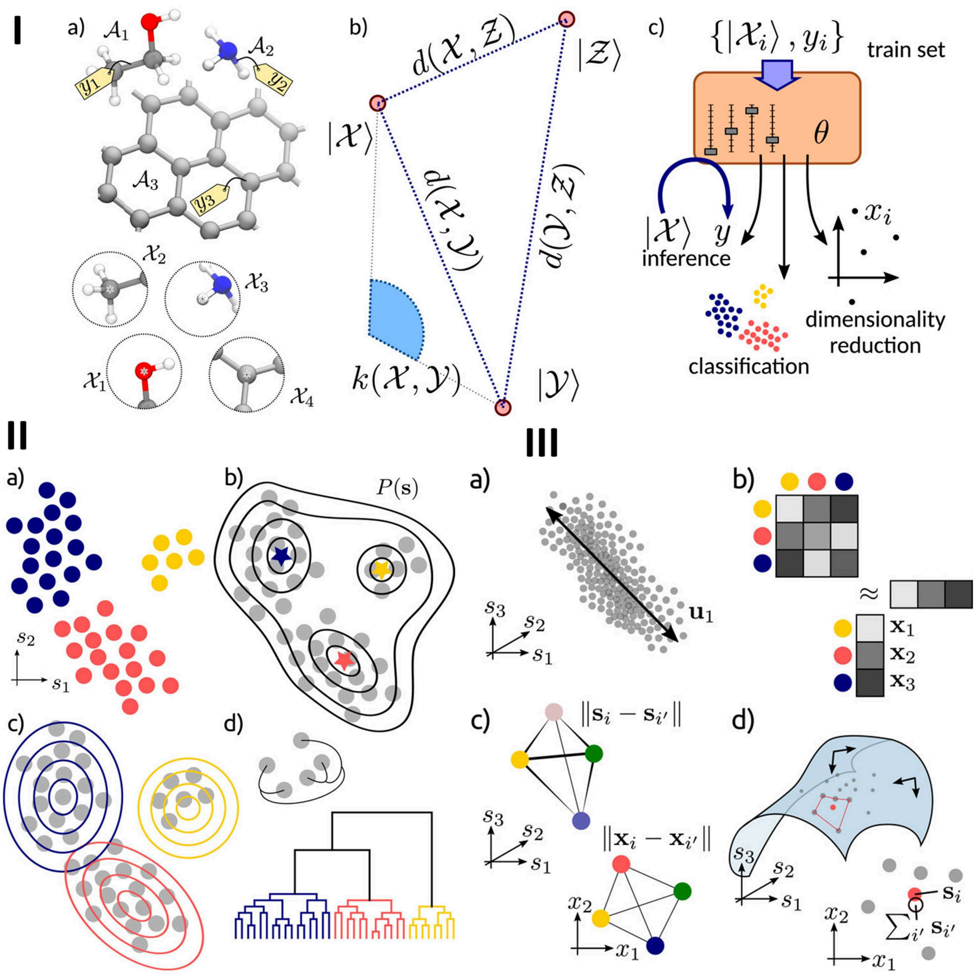


FIGURE 6 | (I) Schematic representation of the main components of atomistic ML. **(a)** the inputs of the model are structures *A* or local environments *X*, **(b)** the mathematical representation of the inputs, based on vectors of features |*X* 〉, a measure of similarity d, or a kernel k, **(c)** the ML model, controlled by a series of parameters θ, and trained based on a set of inputs. **(II)** An overview of the clustering methods, including **(a)** a set of data points clustered according to their hidden common features, **(b)** a density-based clustering for identifying maxima in the probability distribution of inputs, **(c)** distribution-based clustering for finding a model of the data distribution based on the combination of clustering probabilities, and **(d)** hierarchical clustering for identifying natural clusters of the inputs. **(III)** Summary of dimensional reduction techniques, including principal component analysis (PCA) for establishing the most relevant subspace retaining the largest fraction of the input data variance, **(b)** a kernel-based method, **(c)** multidimensional scaling for reproducing in low dimension the similarity between high-dimensional data points. Reprinted with permission from Ceriotti (2019).

considering an external validation set of 6,000 compounds showing the potential transferability of the model within in-class compounds. It was notable to outline QM-calculated energies, with a mean-absolute error of ca. 15 kcal/mol, without using the Schrodinger Equation in the ML algorithm. It was also suggested that the DLNs-based model should outperform the traditional ML-approach (Goh et al., 2017).

More recently, an alternative approach based ML algorithms for supplementing existing QM algorithms was proposed (Ramakrishnan et al., 2015). A Δ-learning approach, involving a ML correction term was developed. Such correction was used in DFT calculated properties for predicting the corresponding quantity at the G4MP2 level of theory. This combined QM/ML approach gathers approximate but fast legacy QM approximations and big-data based QM estimates, trained on results across chemical space, despite being applied using only traditional ML algorithms (Ramakrishnan et al., 2015).

Gómez-Bombarelli et al. (2018) have applied DL for generating and optimizing functional compounds, such as drug-like molecules. The proposed model allowed converting discrete representations of molecules from and into a multidimensional continuous representation, and generating new molecules for exploration and optimization.

A DLN was trained on a a large set of existing chemical structures to build an encoder, which converts the discrete representation of a molecule into a continuous vector, a decoder, that transforms the continuous vector into discrete molecular representations (e.g., SMILES string), and a predictor,

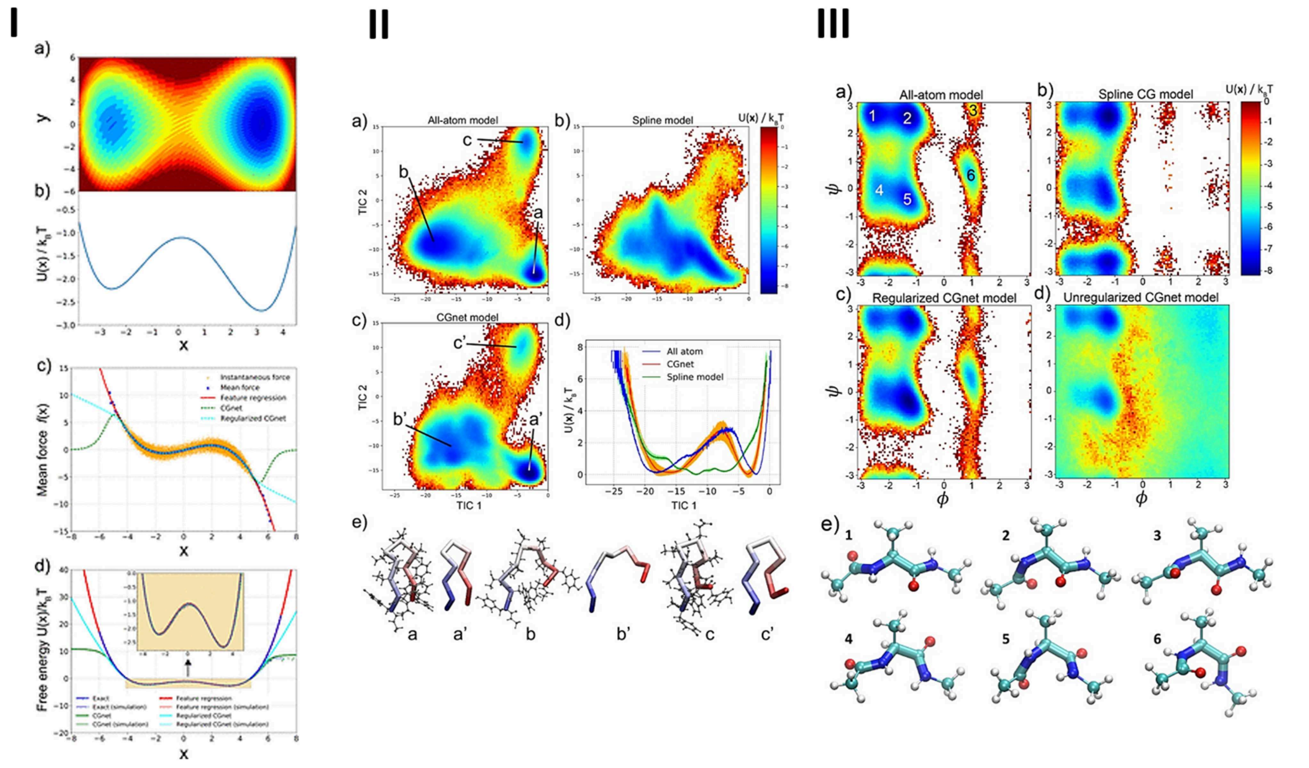


FIGURE 7 | (I) Machine-learned coarse-graining of dynamics in **(a)** a two-dimensional potential, showing the **(b)** exact free-energy along x, comparison of **(c)** the instantaneous forces and the learned mean forces using feature regression and coarse-grained neural network models with the exact forces, and **(d)** the potential-of-mean-force along x, predicted by feature regression, and coarse-grained neural network models with the exact free energy. **(II)** Free-energy profiles and representative structures of alanine dipeptide simulated using all-atom and machine-learned coarse-grained models: **(a)** free-energy reference as a function of the dihedral angles, obtained from the histograms of all-atom simulations, **(b)** standard coarse-grained model using a sum of splines of individual internal coordinates, **(c)** regularized coarse-grained neural network models, **(d)** unregularized networks, **(e)** representative structures extracted from the free-energy minima, from atomistic simulation (ball-and-stick representation) and regularized coarse-grained neural network simulation (licorice representation). **(III)** Free-energy landscape of Chignolin for the different models, obtained from the **(a)** all-atom simulation, as a function of the first two TICA coordinates, **(b)** spline model, as a function of the same two coordinates used in the all-atom model, **(c)** coarse-grained neural network model, as a function of the same two coordinates. **(d)** Comparison of the one-dimensional free-energy profile as a function of the first TICA coordinate, reflecting the folding/unfolding transition, for the all-atom (blue), spline (green), and coarse-grained neural network models (red). **(e)** Representative Chignolin conformations in the three minima from (a–c) all-atom simulation and (a′-c′) coarse-grained neural network model. Reprinted with permission from Wang et al. (2019).

which estimates chemical properties from the latent continuous vector representation of the molecule. These representations allowed generating new chemical structures automatically by employing simple operations in the latent space (e.g., decoding random vectors, perturbing defined chemical structures, and interpolating between molecules), and applying gradient-based optimization for a oriented-search of functional molecules (Gómez-Bombarelli et al., 2018).

DLNs have also been applied for exploring the molecular conformational space of proteins. Some authors (Degiacomi, 2019) have demonstrated that generative neural networks trained on protein structures, extracted from molecular simulation, can be employed to create new conformations complementing pre-existing ones. The model was trained and tested in a protein-protein docking scenario to account for specific motions occurring upon binding.

The fewer examples of DLNs applications in quantum chemistry suggest that it is still in an earlier stage of development compared to other approaches including computational structural biology and computer-aided drug design.

Planning and Predicting Reactions and Routes

Some practical questions in organic chemistry have been addressed by ML approaches, including the identification of the most suitable synthesis method for a specific compound and the optimal conditions (reactants, solvent, catalyst, temperature, and among others) for ensuring region/chemo/stereo selectivity and obtaining the highest yields, estimating the precise rate, yield and time for the reaction, predicting major/minor product, and also evaluating similarity between reactions (Wei et al., 2016; Ahneman et al., 2018).

Making predictions in reactive chemical systems can also resort to DL. Segler and Waller (2017) and Segler et al. (2018) have predicted reaction rules considering fundamental substructures of reactants and products, allowing to return a product, given a reactant as input, and vice versa. In simple terms, a reaction rule is a pattern guiding the interaction process for a set of reactants and suggesting potential chemical products. As the knowledge available in often inaccurate, such rules are often ambiguous or even incomplete (Kishimoto et al., 2018). However, there are some successful examples, such as the recent outcomes of Chematica. Grzybowski et al. (2018) have assembled the relevant transformations that connect chemical species into a large network. The latter have codified and organized the known pathways through chemical space and displays nodes of molecules, elements and chemical reactions, collected by linking reactants to products on the basis of core reactions.

The Chematica platform comprises network theory, high-performance computing, artificial intelligence, and expert chemical knowledge to accelerate the design of synthetic pathways leading to new targets. However, the experimental verification of the respective predictions was carried out recently (Grzybowski et al., 2018). The authors have described the results of a systematic approach in which synthetic pathways leading to eight targets with distinct structures and of medicinal relevance were designed without human supervision and experimentally validated. There are other prominent products such as ChemPlanner, and Synthia created from databases of rules for chemical transformations. Both platforms incorporate ML algorithms and allows navigating through chemical space using those rules and suggesting to the user possible ways to synthesize a target molecule. Synthia also employs MD, quantum mechanics, and electronic properties to infer on the viability of a transformation and on the stability of an intermediate along a synthesis route (Klucznik et al., 2018).

Reaction prediction and retrosynthesis are the mainstays of organic chemistry. Retrosynthesis has been used for planning synthesis of small organic molecules, in which target molecules are recursively converted into progressively simpler precursors (Segler and Waller, 2017). However, the results obtained from the *in silico* version of this process are not, in general, adequate. Rule-based procedures have been extensively employed for solving, computationally, both reaction prediction and retrosynthesis. However, reactivity conflicts are often generated, since reaction rules tend to ignore the molecular context. It is often difficult to predict how a compound would behave in practice, unless an experiment is carried out (Granda et al., 2018). Evaluating a candidate sequence of reaction steps means that the synthesis of a given compound is also difficult. In chemical synthesis planning, Szymkuć et al. (2016) have discussed these issues. Segler and Waller have reported (Segler et al., 2018) that the prioritization of the most suitable conversion rules, as well as the approach to conflicting or complexity raising issues can be achieved by learning with DLNs. The authors have trained their model on ca. three million reactions, exhibiting accuracies of 97 and 95% for reaction prediction and retrosynthesis, respectively, on a validation set of ca. one million reactions. Following this procedure, the same authors have applied Monte Carlo tree search and symbolic artificial intelligence to find retrosynthetic routes. DLNs were trained on the whole set of published organic reactions (Segler et al., 2018).

Coley et al. (2017, 2018b) have performed DL with features based on the alterations of reactants and have determined scores for putative products. The product was modeled as a true target molecule (product) if it was generated by a reaction covered by the patent literature, and as a false product otherwise. More recently Coley et al. (2018b) have put forward a new definition addressing the synthetic complexity in order to compare with the expected number of reaction steps required for producing target molecules, with known compounds as reasonable starting materials. Specifically, a neural network model was trained on 12 million reactions from the Reaxys database, imposing a pairwise inequality constraint and showing that the products of published chemical reaction are, on average, more synthetically complex than their corresponding reactants.

A graph-link-prediction-based procedure was formulated by Savage et al. (2017) to predict candidate molecules (reactants), given a target molecule (product) as input and to discover adequate synthesis routes for producing the targets. This was employed over the Network of Organic Chemistry constructed from eight million chemical reactions described in the US patent literature in the 1976–2013 period (Savage et al., 2017). The proposed evaluation demonstrated that Factorization Machines, trained with chemistry-specific information, outperforms similarity-based methods of chemical structures. In these approaches, a fingerprint is built from a graphical representation of the molecule, containing the respective structural information and chemical features. The latter can be selected using different approaches (Morgan, 1965; Rogers and Hahn, 2010). Some neural graph fingerprints have displayed significant predictive performance (Duvenaud et al., 2015). The detection of molecular active substructures (e.g., a moiety impacting on a disease and a moiety that confers structural stability) can also be performed with ML (Duvenaud et al., 2015).

Researchers have also designed a chemical-handling robot for screening and predicting chemical reactivity using ML. The authors have found four novel reactions, demonstrating the respective potential in discovering reactions. Chemical reactions related to many different pathways can lead to a desired molecule. To find the best pathways, discovering new chemical reactivity is crucial to make the processes that produce chemicals, pharmaceuticals and materials more sustainable, environmentally-friendly and efficient. However, discovering new reactions is usually an unpredictable and time-consuming process that's constrained by a top-down approach involving expert knowledge to target a particular molecule.

Other researchers (Granda et al., 2018) have created an organic synthesis robotic ML system able to explore the reactivity several reagents from the bottom-up with no specific target. By performing ca. 10% of 969 possible reactions from a set of 18 reagents, the proposed system allowed predicting the reactivity of the remaining 90% of reactions with an accuracy of 86%. The database was continuously updated by performing multiple experiments based on the reactivity data collected. This allowed discovering new reactions that were inspected to isolate and characterize the new compounds (Granda et al., 2018).

Supporting Analytical Chemistry and Catalysis

Analytical chemistry is possibly the area corresponding to the longest history, but also one that mostly displays embryonic applications of ML. A large number of statistical analyses and ML expert systems have been implemented in analytical chemistry for a long time (e.g., comparing and classifying mass spectra, NMR, or IR through assessments on available compounds) (Lipkowitz and Boyd, 1995; Mayer and Baeumner, 2019). Until recently, ML approaches were mainly employed to explain chemical reactions and to provide valuable predictive insights. Currently, it is possible to predict unexpected reactive outcomes, or relevant mechanistic insights for catalytic processes. A survey of some of these contributions can be found in Durand and Fey (2019).

Other groups (Ghosh et al., 2019) have proposed DL methods for predicting molecular excitation spectra. Considering the electronic density of the states of 132 k organic compounds, the authors have built three different neural network architectures: a multilayer perceptron (MLP), a convolutional neural network (CNN), and a DLNs. The coordinates and charge of the atoms in each molecule were used as inputs for the neural networks. The DLNs reached the best performance with a root-mean-square error (RMSE) of 0.19 eV, while MLP and CNN were able to learn spectra with a RMSE of 0.3 and 0.23 eV, respectively. Both CNN and DLNs allowed identifying subtle variations in the spectral shape. The structures of 10 k organic molecules previously unseen were scanned and the instant predictions on spectra were obtained to identify molecules for further applications (Ghosh et al., 2019).

A new computational approach, denoted as quantitative profile-profile relationship (QPPR) modeling, and based on ML techniques, has been proposed for predicting the pre-discharge chemical profiles of ammunition components from the components of the respective post-discharge gunshot residue (Gallidabino et al., 2019). The predicted profiles can be compared with other measured profiles to perform evidential associations in forensic investigations. Specifically, the approach was optimized and assessed for the prediction of GC-MS profiles of smokeless powders (SLPs) obtained from organic gunshot residues, considering nine ammunition types. A high degree of similarity between predicted and experimentally measured profiles was found, after applying 14 ML techniques, with a median correlation of 0.982 (Gallidabino et al., 2019). Receiver operating characteristic (ROC) analysis was employed to assess association performances, and allowed comparing predicted–predicted and predicted–measured profiles, producing areas under the curve (AUCs) of 0.976 and 0.824, respectively, in extrapolation mode. On the other hand, AUCs of 0.962 and 0.894 were obtained in the interpolation mode. These results were approximated to the values obtained from the comparison of the measured SLP profiles (AUC = 0.998), demonstrating excellent potential to correctly associate evidence in a number of different forensic situations (Gallidabino et al., 2019). The advantages of this approach are numerous and may be extended to other fields in analytical sciences that routinely experience mutable chemical signatures, including the analysis of explosive devices, toxicological samples and environmental pollutants (Gallidabino et al., 2019).

The integration of ML-based algorithms in a chemosensor has also pointed out the future of ML and the artificial internet of things applicability, i.e., optimized sensors, linked to a central data analysis unit via wireless (Mayer and Baeumner, 2019).

Additionally, researchers have used ML to develop tools for predicting catalytic components and dynamics. For instance, the identification and prediction of ligands for metal-catalyzed coupling reaction have been conducted for designing a synthetic economic and ecological route, with the potential to be expanded into a system of pharmaceutical interest (Durand and Fey, 2019). Durand and Fey have recently summarized calculations of several ligand descriptors, focusing on homogeneous organometallic catalysis. Different approaches for calculating steric and electronic parameters were also reviewed and assessed, and a set of descriptors for a wide range of ligands (e.g., 30 monodentate phosphorus (III) donor ligands, 23 bidentate P,P-donor ligands, and 30 carbenes) were collected.

Different case studies covering the application of these descriptors, including maps and models and DFT calculations, have been discussed, demonstrating the usefulness of descriptor-oriented studies of catalysis for guiding experiments and successfully evaluate and compare the proposed models (Durand and Fey, 2019).

Li and Eastgate (2019) have designed a ML-based tool for acting on transition metal-catalyzed carbon–nitrogen coupling reactions encompassing phosphine ligands, which are often involved in pharmaceutical syntheses. The data set of the system was composed of literature documents reporting coupling reactions with phosphine ligands. The input variables were the molecular features of ligand electrophiles and nucleophiles, and the phosphine ligands were de output obtained in successful reactions. The tools used substrate fingerprints, to build a multiclass predictive model and identify the ligands prone to function in a Pd-catalyzed C–N coupling reaction. The resulting probabilities were associated to the corresponding ligand (cPMIs) to determine a probability-weighted predicted holistic PMI for the transformation, considering the synthesis of the ligand. This novel ML approach were developed for estimating the probability of success for ligands, given specified electrophile and nucleophile combinations, illustrated in the a Pd-catalyzed C–N coupling context. The neural network allowed thus improving the predictive performance of the top-N accuracy over other ML approaches. Further application of this tool will foster the development of frameworks based on criteria-decision analytics, optimizing the design of manufacturing processes.

Designing catalysts using computational approaches is also a major challenge in chemistry. Conventional approaches have been restricted to calculate properties for a complex and large number of potential catalysts. More recently, innovative approaches for inverse design have emerged, for finding the desired property and optimizing the respective chemical structure. The chemical space has been explored by combining gradient-based optimization, alchemical transformations, and

ML. These efforts have been duly reviewed in the context of inverse design and relevance to developing catalytic technologies (Freeze et al., 2019). These approaches have offered new opportunities for identifying catalysts using efficient methods that circumvent the need for high-throughput screening and reduce the array of compounds and materials displaying the target properties and can be experimentally validated. For instance, inverse design can be employed for modulating catalytic activity via alterations in the first and second coordination spheres of the catalyst binding site (e.g., functionality of catalytic cofactors in enzymes).

One possible approach to inverse design is to use the synthetic accessibility score, commonly used for drug molecules, in the scoring functions of inverse design for ensuring synthetic feasibility. For that purpose, empirical parameters can be used to describe molecules without the cost of using 3D coordinates for an entire structure and without using a model to describe the complex interactions from geometries.

The major progress on inverse design relies on optimization algorithms, which govern the process for exploring a specific space, improving identification rates of parameters that allows optimizing the value of the scoring function. For example, the Classical Optimal Control Optimization algorithm, used for global energy minimization, is based on the diffeomorphic modulation under the observable-response-preserving homotopy algorithm, and lead the classical dynamics of a probe particle, driven by an external field for reaching the optimal value of a multidimensional function, by adjusting iteratively field control parameters over the gradient of the scoring function related to the controls. However, the respective use for scoring functions in inverse design applications still remain a challenge (Freeze et al., 2019). Scoring functions allow correlating molecular descriptors to catalytic properties for finding catalysts via gradient-based optimization. In a simple example, similar molecules often display distinct catalytic activity due to subtle effects that must be detected by scoring functions. Such effects may be determined by combining experimentation to build adequate training sets of systems with different values of selected properties for determining feature sets able to detect such properties. ML can also be used to evaluate performance scores for GA-based methods.

The application of autoencoders have allowed transforming SMILES representations of compounds into a continuous latent space in order to optimize chemical properties, including synthetic accessibility score and Quantitative Estimation of Drug Likeness. Additionally, by resorting to gradient-based methods the latent space can be intersected to predict new candidate structures for being synthesized and tested.

The integration of inverse design, gradient-based optimization and ML is a very promising strategy to shorten the long path toward catalyst discovery (Freeze et al., 2019). Also, other related methods that have been implemented to scrutinize the chemical space for drug development can be applied for catalyst discovery, as described in Freeze et al. (2019).

CONCLUDING REMARKS

This review has sought to provide a sample of ML approaches that support the major research trends in Chemistry, especially in computational chemistry, focusing on DLNs. Such an approaches have offered the possibility of solving chemical problems that cannot be described and explained via conventional methods. In the last few years, the application of ML to the optimization and prediction of molecular properties has become very popular, since more researchers are trained and acquired technical skills to develop and use such methods. ML applications are area-dependent and follow, in fact, a more or less obvious pattern. For instance, medicinal chemistry excels in structure-activity relationships. In other words, each sub-field is progressing essentially in activities that belong to its core subjects. It seems that these fields are evolving naturally, and we cannot identify significant disruptive trends.

Despite the historical route of ML methods involving the implementation of clustering or dimensionality reduction approaches, to provide a simple, low dimensional, or coarse-grained representations of structural and dynamical patterns of complex chemical systems, the interplay between innovative ML-driven predictions and molecular simulations can be combined to make time-consuming electronic calculations feasible, obtain accurate interatomic potentials on complex systems, and provide a rational design of molecules and materials. However, the convergence between different ML algorithms is a major challenge to achieve a definite progress in the chemistry fields.

Unsupervised learning may also contribute to elucidate the operating aspects of supervised algorithms, while supervised approaches may contribute to the formulation of objective metrics to evaluate the performance of unsupervised approaches.

In Chemistry DL is still at an incipient stage, particularly in computational chemistry, although the pace of contributions has been increasing very recently. One of the major challenges is the disparity, quality and interpretability of the generated outcomes. Paired with the sophistication and ability of GPU-accelerated computing for training DLNs and the massive growth of chemical information used for training DLNs, it is anticipated that DL algorithms will be an invaluable engine for computational chemistry. DL uses a hierarchical cascade of non-linear functions allowing to learn representations and capture the required features from raw chemical data, necessary for predicting target physicochemical properties.

Considering the recent effort on the topic, DL models have been implemented in various Chemistry sub-fields, including quantum-chemistry, compound and materials design, with superior performances to conventional ML algorithms. There is still tremendous room for improved model accuracy and interpretability. While industrial sectors will continue driving many advances, academia will continue playing a critical role in supplying innovative technical and practical contributions, as well as in fostering cross-disciplinary cooperation.

AUTHOR CONTRIBUTIONS

TC performed the bibliometrics analysis, collected the relevant studies in the context of the review, structured, and wrote the paper. AP directed the work, contributed to the interpretation of the data, and to the structure of the review. Both authors reviewed the manuscript.

FUNDING

The Coimbra Chemistry Centre (CQC) was supported by the Portuguese Agency for Scientific Research, Fundação para a Ciência e a Tecnologia (FCT), through Project UID/QUI/00313/2019.

REFERENCES

Agrafiotis, D. K., Cedeño, W., and Lobanov, V. S. (2002). On the use of neural network ensembles in QSAR and QSPR. *J. Chem. Inf. Comput. Sci.* 42, 903–911. doi: 10.1021/ci0203702

Ahn, S., Hong, M., Sundararajan, M., Ess, D. H., and Baik, M.-H. (2019). Design and optimization of catalysts based on mechanistic insights derived from quantum chemical reaction modeling. *Chem. Rev.* 119, 6509–6560. doi: 10.1021/acs.chemrev.9b00073

Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D., and Doyle, A. G. (2018). Predicting reaction performance in C–N cross-coupling using machine learning. *Science* 360, 186–190. doi: 10.1126/science.aar5169

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* 33, 831–838. doi: 10.1038/nbt.3300

Artrith, N., and Urban, A. (2016). An implementation of artificial neural-network potentials for atomistic materials simulations: performance for TiO_2. *Comput. Mater. Sci.* 114, 135–150. doi: 10.1016/j.commatsci.2015.11.047

Arús-Pous, J., Blaschke, T., Ulander, S., Reymond, J.-L., Chen, H., and Engkvist, O. (2019). Exploring the GDB-13 chemical space using deep generative models. *J. Cheminform.* 11:20. doi: 10.1186/s13321-019-0341-z

Aspuru-Guzik, A., Baik, M.-H., Balasubramanian, S., Banerjee, R., Bart, S., Borduas-Dedekind, N., et al. (2019). Charting a course for chemistry. *Nat. Chem.* 11, 286–294. doi: 10.1038/s41557-019-0236-7

Bartók, A. P., Kondor, R., and Csányi, G. (2013). On representing chemical environments. *Phys. Rev. B* 87:184115. doi: 10.1103/PhysRevB.87.184115

Bartók, A. P., Payne, M. C., Kondor, R., and Csányi, G. (2010). Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* 104:136403. doi: 10.1103/PhysRevLett.104.136403

Behler, J. (2011a). Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* 134:074106. doi: 10.1063/1.3553717

Behler, J. (2011b). Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. *Phys. Chem. Chem. Phys.* 13, 17930–17955. doi: 10.1039/c1cp21668f

Behler, J. (2016). Perspective: machine learning potentials for atomistic simulations. *J. Chem. Phys.* 145:170901. doi: 10.1063/1.4966192

Behler, J., and Parrinello, M. (2007). Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* 98:146401. doi: 10.1103/PhysRevLett.98.146401

Benjamin, S.-L., Carlos, O., Gabriel, L., G., and Alan, A.-G. (2017). Optimizing Distributions Over Molecular Space. An Objective-Reinforced Generative Adversarial Network for Inverse-design Chemistry (ORGANIC). *ChemRxiv [Preprint]*. doi: 10.26434/chemrxiv.5309668.v3

Bose, S., Dhawan, D., Nandi, S., Sarkar, R. R., and Ghosh, D. (2018). Machine learning prediction of interaction energies in rigid water clusters. *Phys. Chem. Chem. Phys.* 20, 22987–22996. doi: 10.1039/C8CP03138J

Botu, V., Batra, R., Chapman, J., and Ramprasad, R. (2017). Machine learning force fields: construction, validation, and outlook. *J. Phys. Chem. C* 121, 511–522. doi: 10.1021/acs.jpcc.6b10908

Botu, V., and Ramprasad, R. (2015a). Adaptive machine learning framework to accelerate ab initio molecular dynamics. *Int. J. Quantum Chem.* 115, 1074–1083. doi: 10.1002/qua.24836

Botu, V., and Ramprasad, R. (2015b). Learning scheme to predict atomic forces and accelerate materials simulations. *Phys. Rev. B* 92:094306. doi: 10.1103/PhysRevB.92.094306

Brockherde, F., Vogt, L., Li, L., Tuckerman, M. E., Burke, K., and Müller, K.-R. (2017). Bypassing the Kohn-Sham equations with machine learning. *Nat. Commun.* 8:872. doi: 10.1038/s41467-017-00839-3

Brown, N., Fiscato, M., Segler, M. H. S., and Vaucher, A. C. (2019). GuacaMol: benchmarking models for *de novo* molecular design. *J. Chem. Inf. Model.* 59, 1096–1108. doi: 10.1021/acs.jcim.8b00839

Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., and Walsh, A. (2018). Machine learning for molecular and materials science. *Nature* 559, 547–555. doi: 10.1038/s41586-018-0337-2

Ceriotti, M. (2019). Unsupervised machine learning in atomistic simulations, between predictions and understanding. *J. Chem. Phys.* 150:150901. doi: 10.1063/1.5091842

Chakravarti, S. K. (2018). Distributed representation of chemical fragments. *ACS Omega* 3, 2825–2836. doi: 10.1021/acsomega.7b02045

Chan, H., Cherukara, M. J., Narayanan, B., Loeffler, T. D., Benmore, C., Gray, S. K., et al. (2019). Machine learning coarse grained models for water. *Nat. Commun.* 10:379. doi: 10.1038/s41467-018-08222-6

Chandrasekaran, A., Kamal, D., Batra, R., Kim, C., Chen, L., and Ramprasad, R. (2019). Solving the electronic structure problem with machine learning. *NPJ Comput. Mater.* 5:22. doi: 10.1038/s41524-019-0162-7

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., and Blaschke, T. (2018). The rise of deep learning in drug discovery. *Drug Discov. Today* 23, 1241–1250. doi: 10.1016/j.drudis.2018.01.039

Chen, M., Yu, T.-Q., and Tuckerman, M. E. (2015). Locating landmarks on high-dimensional free energy surfaces. *Proc. Natl. Acad. Sci. U.S.A.* 112:3235. doi: 10.1073/pnas.1418241112

Chmiela, S., Sauceda, H. E., Müller, K.-R., and Tkatchenko, A. (2018). Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* 9, 3887–3887. doi: 10.1038/s41467-018-06169-2

Chmiela, S., Tkatchenko, A., Sauceda, H. E., Poltavsky, I., Schütt, K. T., and Müller, K.-R. (2017). Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* 3:e1603015. doi: 10.1126/sciadv.1603015

Choi, H., Kang, H., Chung, K.-C., and Park, H. (2019). Development and application of a comprehensive machine learning program for predicting molecular biochemical and pharmacological properties. *Phys. Chem. Chem. Phys.* 21, 5189–5199. doi: 10.1039/C8CP07002D

Christensen, A. S., Faber, F. A., and von Lilienfeld, O.A. (2019). Operators in quantum machine learning: Response properties in chemical space. *J. Chem. Phys.* 150:064105. doi: 10.1063/1.5053562

Chuang, K. V., and Keiser, M. J. (2018a). Adversarial controls for scientific machine learning. *ACS Chem. Biol.* 13, 2819–2821. doi: 10.1021/acschembio.8b00881

Chuang, K. V., and Keiser, M. J. (2018b). Comment on "predicting reaction performance in C–N cross-coupling using machine learning". *Science* 362:eaat8603. doi: 10.1126/science.aat8603

Coley, C. W., Barzilay, R., Jaakkola, T. S., Green, W. H., and Jensen, K. F. (2017). Prediction of organic reaction outcomes using machine learning. *ACS Central Sci.* 3, 434–443. doi: 10.1021/acscentsci.7b00064

Coley, C. W., Green, W. H., and Jensen, K. F. (2018a). Machine learning in computer-aided synthesis planning. *Acc. Chem. Res.* 51, 1281–1289. doi: 10.1021/acs.accounts.8b00087

Coley, C. W., Rogers, L., Green, W. H., and Jensen, K. F. (2018b). SCScore: synthetic complexity learned from a reaction corpus. *J. Chem. Inf. Model.* 58, 252–261. doi: 10.1021/acs.jcim.7b00622

Cortés-Borda, D., Kutonova, K. V., Jamet, C., Trusova, M. E., Zammattio, F., Truchet, C., et al. (2016). Optimizing the Heck–Matsuda reaction in flow with a constraint-adapted direct search algorithm. *Organ. Process Res. Dev.* 20, 1979–1987. doi: 10.1021/acs.oprd.6b00310

Coveney Peter, V., Dougherty Edward, R., and Highfield Roger, R. (2016). Big data need big theory too. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* 374:20160153. doi: 10.1098/rsta.2016.0153

Curtarolo, S., Hart, G. L. W., Nardelli, M. B., Mingo, N., Sanvito, S., and Levy, O. (2013). The high-throughput highway to computational materials design. *Nat. Mater.* 12, 191–201. doi: 10.1038/nmat3568

De, S., Bartók, A. P., Csányi, G., and Ceriotti, M. (2016). Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* 18, 13754–13769. doi: 10.1039/C6CP00415F

Degiacomi, M. T. (2019). Coupling molecular dynamics and deep learning to mine protein conformational space. *Structure* 27, 1034–1040.e1033. doi: 10.1016/j.str.2019.03.018

Dral, P. O., Von Lilienfeld, O. A., and Thiel, W. (2015). Machine learning of parameters for accurate semiempirical quantum chemical calculations. *J. Chem. Theory Comput.* 11, 2120–2125. doi: 10.1021/acs.jctc.5b00141

Durand, D. J., and Fey, N. (2019). Computational ligand descriptors for catalyst design. *Chem. Rev.* 119, 6561–6594. doi: 10.1021/acs.chemrev.8b00588

Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., et al. (2015). "Convolutional networks on graphs for learning molecular fingerprints," in *Advances in Neural Information Processing Systems*, eds C. Cortes, N. D. Lawrence, D. D Lee, M. Sugiyama, R. Garnett (Montreal, QC), 2224–2232.

Ekins, S. (2016). The next Era: deep learning in pharmaceutical research. *Pharm. Res.* 33, 2594–2603. doi: 10.1007/s11095-016-2029-7

El-Atta, A. H. A., and Hassanien, A. E. (2017). Two-class support vector machine with new kernel function based on paths of features for predicting chemical activity. *Inf. Sci.* 403–404, 42–54. doi: 10.1016/j.ins.2017.04.003

Elton, D. C., Boukouvalas, Z., Butrico, M. S., Fuge, M. D., and Chung, P. W. (2018). Applying machine learning techniques to predict the properties of energetic materials. *Sci. Rep.* 8:9059. doi: 10.1038/s41598-018-27344-x

Elton, D. C., Boukouvalas, Z., Fuge, M. D., and Chung, P. W. (2019). Deep learning for molecular design—a review of the state of the art. *Mol. Syst. Design Eng.* 4, 828–849. doi: 10.1039/C9ME00039A

Estrada, J. G., Ahneman, D. T., Sheridan, R. P., Dreher, S. D., and Doyle, A. G. (2018). Response to comment on "predicting reaction performance in C–N cross-coupling using machine learning". *Science* 362:eaat8763. doi: 10.1126/science.aat8763

Faber, F. A., Hutchison, L., Huang, B., Gilmer, J., Schoenholz, S. S., Dahl, G. E., et al. (2017). Prediction errors of molecular machine learning models lower than hybrid DFT error. *J. Chem. Theory Comput.* 13, 5255–5264. doi: 10.1021/acs.jctc.7b00577

Faber, F. A., Lindmaa, A., Von Lilienfeld, O. A., and Armiento, R. (2016). Machine learning energies of 2 million elpasolite (ABC$_2$D$_6$) crystals. *Phys. Rev. Lett.* 117:135502. doi: 10.1103/PhysRevLett.117.135502

Fleming, N. (2018). How artificial intelligence is changing drug discovery. *Nature* 557, S55–S55. doi: 10.1038/d41586-018-05267-x

Freeze, J. G., Kelly, H. R., and Batista, V. S. (2019). Search for catalysts by inverse design: artificial intelligence, mountain climbers, and alchemists. *Chem. Rev.* 119, 6595–6612. doi: 10.1021/acs.chemrev.8b00759

Fuchs, J.-A., Grisoni, F., Kossenjans, M., Hiss, J. A., and Schneider, G. (2018). Lipophilicity prediction of peptides and peptide derivatives by consensus machine learning. *Medchemcomm* 9, 1538–1546. doi: 10.1039/C8MD00370J

Gallidabino, M. D., Barron, L. P., Weyermann, C., and Romolo, F. S. (2019). Quantitative profile–profile relationship (QPPR) modelling: a novel machine learning approach to predict and associate chemical characteristics of unspent ammunition from gunshot residue (GSR). *Analyst* 144, 1128–1139. doi: 10.1039/C8AN01841C

Gao, H., Struble, T. J., Coley, C. W., Wang, Y., Green, W. H., and Jensen, K. F. (2018). Using machine learning to predict suitable conditions for organic reactions. *ACS Central Sci.* 4, 1465–1476. doi: 10.1021/acscentsci.8b00357

Gasteiger, J., and Zupan, J. (1993). Neural networks in chemistry. *Angew. Chem. Int. Ed. Eng.* 32, 503–527. doi: 10.1002/anie.199305031

Ghosh, K., Stuke, A., Todorović, M., Jørgensen, P. B., Schmidt, M. N., Vehtari, A., et al. (2019). Deep learning spectroscopy: neural networks for molecular excitation spectra. *Adv. Sci.* 6:1801367. doi: 10.1002/advs.201801367

Goh, G. B., Hodas, N. O., and Vishnu, A. (2017). Deep learning for computational chemistry. *J. Comput. Chem.* 38, 1291–1307. doi: 10.1002/jcc.24764

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., et al. (2018). Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Sci.* 4, 268–276. doi: 10.1021/acscentsci.7b00572

Granda, J. M., Donina, L., Dragone, V., Long, D.-L., and Cronin, L. (2018). Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* 559, 377–381. doi: 10.1038/s41586-018-0307-8

Green, C. P., Engkvist, O., and Pairaudeau, G. (2018). The convergence of artificial intelligence and chemistry for improved drug discovery. *Future Med. Chem.* 10, 2573–2576. doi: 10.4155/fmc-2018-0161

Griffen, E. J., Dossetter, A. G., Leach, A. G., and Montague, S. (2018). Can we accelerate medicinal chemistry by augmenting the chemist with Big Data and artificial intelligence? *Drug Discov. Today* 23, 1373–1384. doi: 10.1016/j.drudis.2018.03.011

Grisoni, F., Neuhaus, C. S., Gabernet, G., Müller, A. T., Hiss, J. A., and Schneider, G. (2018). Designing anticancer peptides by constructive machine learning. *ChemMedChem* 13, 1300–1302. doi: 10.1002/cmdc.201800204

Gromski, P. S., Henson, A. B., Granda, J. M., and Cronin, L. (2019). How to explore chemical space using algorithms and automation. *Nat. Rev. Chem.* 3, 119–128. doi: 10.1038/s41570-018-0066-y

Grzybowski, B. A., Szymkuć, S., Gajewska, E. P., Molga, K., Dittwald, P., Wołos, A., et al. (2018). Chematica: a story of computer code that started to think like a chemist. *Chem* 4, 390–398. doi: 10.1016/j.chempr.2018.02.024

Gupta, A., Müller, A. T., Huisman, B. J. H., Fuchs, J. A., Schneider, P., and Schneider, G. (2018). Generative recurrent networks for *de novo* drug design. *Mol. Inform.* 37:1700111. doi: 10.1002/minf.201700111

Hansen, K., Biegler, F., Ramakrishnan, R., Pronobis, W., Von Lilienfeld, O. A., Müller, K.-R., et al. (2015). Machine learning predictions of molecular properties: accurate many-body potentials and nonlocality in chemical space. *J. Phys. Chem. Lett.* 6, 2326–2331. doi: 10.1021/acs.jpclett.5b00831

Hansen, K., Montavon, G., Biegler, F., Fazli, S., Rupp, M., Scheffler, M., et al. (2013). Assessment and validation of machine learning methods for predicting molecular atomization energies. *J. Chem. Theory Comput.* 9, 3404–3419. doi: 10.1021/ct400195d

Harel, S., and Radinsky, K. (2018). Prototype-based compound discovery using deep generative models. *Mol. Pharm.* 15, 4406–4416. doi: 10.1021/acs.molpharmaceut.8b00474

Häse, F., Fdez. Galván, I., Aspuru-Guzik, A., Lindh, R., and Vacher, M. (2019). How machine learning can assist the interpretation of ab initio molecular dynamics simulations and conceptual understanding of chemistry. *Chem. Sci.* 10, 2298–2307. doi: 10.1039/C8SC04516J

Häse, F., Roch, L. M., and Aspuru-Guzik, A. (2018). Chimera: enabling hierarchy based multi-objective optimization for self-driving laboratories. *Chem. Sci.* 9, 7642–7655. doi: 10.1039/C8SC02239A

He, Y., Cubuk, E. D., Allendorf, M. D., and Reed, E. J. (2018). Metallic metal–organic frameworks predicted by the combination of machine learning methods and Ab initio calculations. *J. Phys. Chem. Lett.* 9, 4562–4569. doi: 10.1021/acs.jpclett.8b01707

Hegde, G., and Bowen, R. C. (2017). Machine-learned approximations to density functional theory hamiltonians. *Sci. Rep.* 7:42669. doi: 10.1038/srep42669

Hiller, S. A., Golender, V. E., Rosenblit, A. B., Rastrigin, L. A., and Glaz, A. B. (1973). Cybernetic methods of drug design. I. Statement of the problem—the perceptron approach. *Comput. Biomed. Res.* 6, 411–421. doi: 10.1016/0010-4809(73)90074-8

Houben, C., and Lapkin, A. A. (2015). Automatic discovery and optimization of chemical processes. *Curr. Opin. Chem. Eng.* 9, 1–7. doi: 10.1016/j.coche.2015.07.001

Houben, C., Peremezhney, N., Zubov, A., Kosek, J., and Lapkin, A. A. (2015). Closed-loop multitarget optimization for discovery of new emulsion polymerization recipes. *Organ. Process Res. Dev.* 19, 1049–1053. doi: 10.1021/acs.oprd.5b00210

Huang, S.-D., Shang, C., Kang, P.-L., and Liu, Z.-P. (2018). Atomic structure of boron resolved using machine learning and global sampling. *Chem. Sci.* 9, 8644–8655. doi: 10.1039/C8SC03427C

Hughes, Z. E., Thacker, J. C. R., Wilson, A. L., and Popelier, P. L. A. (2019). Description of potential energy surfaces of molecules using FFLUX machine learning models. *J. Chem. Theory Comput.* 15, 116–126. doi: 10.1021/acs.jctc.8b00806

Iype, E., and Urolagin, S. (2019). Machine learning model for non-equilibrium structures and energies of simple molecules. *J. Chem. Phys.* 150:024307. doi: 10.1063/1.5054968

Janet, J. P., Chan, L., and Kulik, H. J. (2018). Accelerating chemical discovery with machine learning: simulated evolution of spin crossover complexes with an artificial neural network. *J. Phys. Chem. Lett.* 9, 1064–1071. doi: 10.1021/acs.jpclett.8b00170

Jaquis, B. J., Li, A., Monnier, N. D., Sisk, R. G., Acree, W. E., and Lang, A. S. (2019). Using machine learning to predict enthalpy of solvation. *J. Solution Chem.* 48, 564–573. doi: 10.1007/s10953-019-00867-1

Jensen, J. H. (2019). A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space. *Chem. Sci.* 10, 3567–3572. doi: 10.1039/C8SC05372C

Jha, D., Ward, L., Paul, A., Liao, W.-K., Choudhary, A., Wolverton, C., et al. (2018). ElemNet: deep learning the chemistry of materials from only elemental composition. *Sci. Rep.* 8:17593. doi: 10.1038/s41598-018-35934-y

Jørgensen, P. B., Mesta, M., Shil, S., Lastra, J. M. G., Jacobsen, K. W., Thygesen, K. S., et al. (2018a). Machine learning-based screening of complex molecules for polymer solar cells. *J. Chem. Phys.* 148:241735. doi: 10.1063/1.5023563

Jørgensen, P. B., Schmidt, M. N., and Winther, O. (2018b). Deep generative models for molecular science. *Mol. Inform.* 37:1700133. doi: 10.1002/minf.201700133

Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A., and Zhavoronkov, A. (2017). druGAN: an advanced generative adversarial autoencoder model for *de novo* generation of new molecules with desired molecular properties *in silico*. *Mol. Pharm.* 14, 3098–3104. doi: 10.1021/acs.molpharmaceut.7b00346

Kanamori, K., Toyoura, K., Honda, J., Hattori, K., Seko, A., Karasuyama, M., et al. (2018). Exploring a potential energy surface by machine learning for characterizing atomic transport. *Phys. Rev. B* 97:125124. doi: 10.1103/PhysRevB.97.125124

Kang, S., and Cho, K. (2018). Conditional molecular design with deep generative models. *J. Chem. Inf. Model.* 59, 43–52. doi: 10.1021/acs.jcim.8b00263

Kang, X., Zhao, Y., and Li, J. (2018). Predicting refractive index of ionic liquids based on the extreme learning machine (ELM) intelligence algorithm. *J. Mol. Liq.* 250, 44–49. doi: 10.1016/j.molliq.2017.11.166

Keil, M., Exner, T. E., and Brickmann, J. (2004). Pattern recognition strategies for molecular surfaces: III. Binding site prediction with a neural network. *J. Comput. Chem.* 25, 779–789. doi: 10.1002/jcc.10361

Kishimoto, A., Buesser, B., and Botea, A. (2018). "AI meets chemistry," in *Thirty-Second AAAI Conference on Artificial Intelligence*. Ireland: IBM Research.

Klucznik, T., Mikulak-Klucznik, B., Mccormack, M. P., Lima, H., Szymkuć, S., Bhowmick, M., et al. (2018). Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory. *Chem* 4, 522–532. doi: 10.1016/j.chempr.2018.02.002

Kowalik, M., Gothard, C. M., Drews, A. M., Gothard, N. A., Weckiewicz, A., Fuller, P. E., et al. (2012). Parallel optimization of synthetic pathways within the network of organic chemistry. *Angew. Chem. Int. Ed.* 51, 7928–7932. doi: 10.1002/anie.201202209

Krallinger, M., Rabal, O., Lourenço, A., Oyarzabal, J., and Valencia, A. (2017). Information retrieval and text mining technologies for chemistry. *Chem. Rev.* 117, 7673–7761. doi: 10.1021/acs.chemrev.6b00851

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521:436. doi: 10.1038/nature14539

Lee, A. A., Yang, Q., Bassyouni, A., Butler, C. R., Hou, X., Jenkinson, S., et al. (2019). Ligand biological activity predicted by cleaning positive and negative chemical correlations. *Proc. Natl. Acad. Sci. U.S.A.* 116:3373. doi: 10.1073/pnas.1810847116

Li, H., Collins, C. R., Ribelli, T. G., Matyjaszewski, K., Gordon, G. J., Kowalewski, T., et al. (2018a). Tuning the molecular weight distribution from atom transfer radical polymerization using deep reinforcement learning. *Mol. Syst. Design Eng.* 3, 496–508. doi: 10.1039/C7ME00131B

Li, H., Zhang, Z., and Liu, Z. (2017). Application of artificial neural networks for catalysis: a review. *Catalysts* 7:306. doi: 10.3390/catal7100306

Li, J., and Eastgate, M. D. (2019). Making better decisions during synthetic route design: leveraging prediction to achieve greenness-by-design. *React. Chem. Eng.* 4, 1595–1607. doi: 10.1039/C9RE00019D

Li, Y., Zhang, L., and Liu, Z. (2018b). Multi-objective *de novo* drug design with conditional graph generative model. *J. Cheminform.* 10:33. doi: 10.1186/s13321-018-0287-6

Lipkowitz, K. B., and Boyd, D. B. (1995). *Reviews in Computational Chemistry 6.* New York, NY: Wiley Online Library. doi: 10.1002/9780470125830

Lo, Y.-C., Rensi, S. E., Torng, W., and Altman, R. B. (2018). Machine learning in chemoinformatics and drug discovery. *Drug Discov. Today* 23, 1538–1546. doi: 10.1016/j.drudis.2018.05.010

Mansbach, R. A., and Ferguson, A. L. (2015). Machine learning of single molecule free energy surfaces and the impact of chemistry and environment upon structure and dynamics. *J. Chem. Phys.* 142:105101. doi: 10.1063/1.49 14144

Marques, M. R. G., Wolff, J., Steigemann, C., and Marques, M. A. L. (2019). Neural network force fields for simple metals and semiconductors: construction and application to the calculation of phonons and melting temperatures. *Phys.istry Chem. Phys.* 21, 6506–6516. doi: 10.1039/C8CP05771K

Mater, A. C., and Coote, M. L. (2019). Deep learning in chemistry. *J. Chem. Inf. Model.* 59, 2545–2559. doi: 10.1021/acs.jcim.9b00266

Matsuzaka, Y., and Uesawa, Y. (2019). Optimization of a deep-learning method based on the classification of images generated by parameterized deep snap a novel molecular-image-input technique for quantitative structure-activity relationship (QSAR) analysis. *Front. Bioeng. Biotechnol.* 7, 65–65. doi: 10.3389/fbioe.2019.00065

Mayer, M., and Baeumner, A. J. (2019). A megatrend challenging analytical chemistry: biosensor and chemosensor concepts ready for the internet of things. *Chem. Rev*. 119, 7996–8027. doi: 10.1021/acs.chemrev.8b00719

Merk, D., Grisoni, F., Friedrich, L., and Schneider, G. (2018). Tuning artificial intelligence on the *de novo* design of natural-product-inspired retinoid X receptor modulators. *Commun. Chem.* 1:68. doi: 10.1038/s42004-018-0068-1

Mezei, P. D., and Von Lilienfeld, O. A. (2019). Non-covalent quantum machine learning corrections to density functionals. *arXiv [preprint]. arXiv:*1903.09010.

Microsoft (2018). *Machine Learning, Data Mining and Rethinking Knowledge at KDD 2018.* London, UK: Microsoft.

Miller, T. H., Gallidabino, M. D., Macrae, J. I., Hogstrand, C., Bury, N. R., Barron, L. P., et al. (2018). Machine learning for environmental toxicology: a call for integration and innovation. *Environ. Sci. Technol.* 52, 12953–12955. doi: 10.1021/acs.est.8b05382

Min, K., Choi, B., Park, K., and Cho, E. (2018). Machine learning assisted optimization of electrochemical properties for Ni-rich cathode materials. *Sci. Rep.* 8:15778. doi: 10.1038/s41598-018-34201-4

Mitchell, J. B. O. (2014). Machine learning methods in chemoinformatics. *Wiley interdisciplinary reviews. Comput. Mol. Sci.* 4, 468–481. doi: 10.1002/wcms.1183

Mitchell, T. M. (1997). *Machine Learning*. Burr Ridge, IL: McGraw Hill.

Molga, K., Dittwald, P., and Grzybowski, B. A. (2019). Navigating around patented routes by preserving specific motifs along computer-planned retrosynthetic pathways. *Chem* 5, 460–473. doi: 10.1016/j.chempr.2018.12.004

Montavon, G., Rupp, M., Gobre, V., Vazquez-Mayagoitia, A., Hansen, K., Tkatchenko, A., et al. (2013). Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* 15:095003. doi: 10.1088/1367-2630/15/9/095003

Morawietz, T., Singraber, A., Dellago, C., and Behler, J. (2016). How van der Waals interactions determine the unique properties of water. *Proc. Natl. Acad. Sci. U.S.A.* 113, 8368–8373. doi: 10.1073/pnas.1602375113

Morgan, H. L. (1965). The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *J. Chem. Doc.* 5, 107–113. doi: 10.1021/c160017a018

Müller, A. T., Hiss, J. A., and Schneider, G. (2018). Recurrent neural network model for constructive peptide design. *J. Chem. Inf. Model.* 58, 472–479. doi: 10.1021/acs.jcim.7b00414

Nouira, A., Crivello, J.-C., and Sokolovska, N. (2018). CrystalGAN: learning to discover crystallographic structures with generative adversarial networks. *arXiv [preprint]. arXiv:*1810.11203.

Panteleev, J., Gao, H., and Jia, L. (2018). Recent applications of machine learning in medicinal chemistry. *Bioorgan. Med. Chem. Lett.* 28, 2807–2815. doi: 10.1016/j.bmcl.2018.06.046

Popova, M., Isayev, O., and Tropsha, A. (2018). Deep reinforcement learning for *de novo* drug design. *Sci. Adv.* 4:eaap7885. doi: 10.1126/sciadv.aap7885

Pronobis, W., Schütt, K. T., Tkatchenko, A., and Müller, K.-R. (2018). Capturing intensive and extensive DFT/TDDFT molecular properties with machine learning. *Eur. Phys. J. B* 91:178. doi: 10.1140/epjb/e2018-90148-y

Ramakrishnan, R., Dral, P. O., Rupp, M., and Von Lilienfeld, O. A. (2014). Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* 1:140022. doi: 10.1038/sdata.2014.22

Ramakrishnan, R., Dral, P. O., Rupp, M., and Von Lilienfeld, O. A. (2015). Big data meets quantum chemistry approximations: the Δ-machine learning approach. *J. Chem. Theory Comput.* 11, 2087–2096. doi: 10.1021/acs.jctc.5b00099

Ramakrishnan, R., and Von Lilienfeld, O. A. (2017). "Machine learning, quantum chemistry, and chemical space," in *Reviews in Computational Chemistry,* Vol. 30, eds A. L. Parrill and K. B. Lipkowitz (Wiley), 225–256. doi: 10.1002/9781119356059.ch5

Richmond, C. J., Miras, H. N., De La Oliva, A. R., Zang, H., Sans, V., Paramonov, L., et al. (2012). A flow-system array for the discovery and scale up of inorganic clusters. *Nat. Chem.* 4, 1037–1043. doi: 10.1038/nchem.1489

Rogers, D., and Hahn, M. (2010). Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50, 742–754. doi: 10.1021/ci100050t

Rupp, M. (2015). Machine learning for quantum mechanics in a nutshell. *Int. J. Quantum Chem.* 115, 1058–1073. doi: 10.1002/qua.24954

Rupp, M., Ramakrishnan, R., and Von Lilienfeld, O. A. (2015). Machine learning for quantum mechanical properties of atoms in molecules. *J. Phys. Chem. Lett.* 6, 3309–3313. doi: 10.1021/acs.jpclett.5b01456

Rupp, M., Tkatchenko, A., Müller, K.-R., and Von Lilienfeld, O. A. (2012). Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* 108:058301. doi: 10.1103/PhysRevLett.108.058301

Sadowski, P., Fooshee, D., Subrahmanya, N., and Baldi, P. (2016). Synergies between quantum mechanics and machine learning in reaction prediction. *J. Chem. Inf. Model.* 56, 2125–2128. doi: 10.1021/acs.jcim.6b00351

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.* 3, 210–229. doi: 10.1147/rd.33.0210

Sánchez-Lengeling, B., and Aspuru-Guzik, A. (2017). Learning more, with less. *ACS Central Sci.* 3, 275–277. doi: 10.1021/acscentsci.7b00153

Sanchez-Lengeling, B., and Aspuru-Guzik, A. (2018). Inverse molecular design using machine learning: generative models for matter engineering. *Science* 361, 360–365. doi: 10.1126/science.aat2663

Sanchez-Lengeling, B., Roch, L. M., Perea, J. D., Langner, S., Brabec, C. J., and Aspuru-Guzik, A. (2019). A Bayesian approach to predict solubility parameters. *Adv. Theory Simul.* 2:1800069. doi: 10.1002/adts.201800069

Savage, J., Kishimoto, A., Buesser, B., Diaz-Aviles, E., and Alzate, C. (2017). "Chemical reactant recommendation using a network of organic chemistry," in *Proceedings of the Eleventh ACM Conference on Recommender Systems* (New York, NY: ACM), 210–214.

Schleder, G. R., Padilha, A. C. M., Acosta, C. M., Costa, M., and Fazzio, A. (2019). From DFT to machine learning: recent approaches to materials science–a review. *J. Phys. Mater.* 2:032001. doi: 10.1088/2515-7639/ab084b

Schneider, G. (2018). Generative models for artificially-intelligent molecular design. *Mol. Inform.* 37:1880131. doi: 10.1002/minf.201880131

Schütt, K., Glawe, H., Brockherde, F., Sanna, A., Müller, K., and Gross, E. (2014). How to represent crystal structures for machine learning: towards fast prediction of electronic properties. *Phys. Rev. B* 89:205118. doi: 10.1103/PhysRevB.89.205118

Segler, M. H. S., Preuss, M., and Waller, M. P. (2018). Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* 555, 604–610. doi: 10.1038/nature25978

Segler, M. H. S., and Waller, M. P. (2017). Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chem. A Eur. J.* 23, 5966–5971. doi: 10.1002/chem.201605499

Shen, X., Zhang, T., Broderick, S., and Rajan, K. (2018). Correlative analysis of metal organic framework structures through manifold learning of Hirshfeld surfaces. *Mol. Syst. Design Eng.* 3, 826–838. doi: 10.1039/C8ME00014J

Simões, R. S., Maltarollo, V. G., Oliveira, P. R., and Honorio, K. M. (2018). Transfer and multi-task learning in QSAR modeling: advances and challenges. *Front. Pharmacol.* 9:74. doi: 10.3389/fphar.2018.00074

Smith, C. J., Nikbin, N., Ley, S. V., Lange, H., and Baxendale, I. R. (2011). A fully automated, multistep flow synthesis of 5-amino-4-cyano-1,2,3-triazoles. *Organ. Biomol. Chem.* 9, 1938–1947. doi: 10.1039/c0ob00815j

Smith, J. S., Isayev, O., and Roitberg, A. E. (2017). ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* 8, 3192–3203. doi: 10.1039/C6SC05720A

Smith, J. S., Nebgen, B., Lubbers, N., Isayev, O., and Roitberg, A. E. (2018a). Less is more: Sampling chemical space with active learning. *J. Chem. Phys.* 148:241733. doi: 10.1063/1.5023802

Smith, J. S., Roitberg, A. E., and Isayev, O. (2018b). Transforming computational drug discovery with machine learning and AI. *ACS Med. Chem. Lett.* 9, 1065–1069. doi: 10.1021/acsmedchemlett.8b00437

Snyder, J. C., Rupp, M., Hansen, K., Müller, K.-R., and Burke, K. (2012). Finding density functionals with machine learning. *Phys. Rev. Lett.* 108:253002. doi: 10.1103/PhysRevLett.108.253002

Stein, H. S., Guevarra, D., Newhouse, P. F., Soedarmadji, E., and Gregoire, J. M. (2019a). Machine learning of optical properties of materials – predicting spectra from images and images from spectra. *Chem. Sci.* 10, 47–55. doi: 10.1039/C8SC03077D

Stein, H. S., Soedarmadji, E., Newhouse, P. F., Dan, G., and Gregoire, J. M. (2019b). Synthesis, optical imaging, and absorption spectroscopy data for 179072 metal oxides. *Sci. Data* 6:9. doi: 10.1038/s41597-019-0019-4

Stevens, J. G., Bourne, R. A., Twigg, M. V., and Poliakoff, M. (2010). Real-time product switching using a twin catalyst system for the hydrogenation of furfural in supercritical CO_2. *Angew. Chem. Int. Ed.* 49, 8856–8859. doi: 10.1002/anie.201005092

Szymkuć, S., Gajewska, E. P., Klucznik, T., Molga, K., Dittwald, P., Startek, M., et al. (2016). Computer-assisted synthetic planning: the end of the beginning. *Angew. Chem. Int. Ed.* 55:5904. doi: 10.1002/anie.201506101

Thomsen, J. U., and Meyer, B. (1989). Pattern recognition of the 1H NMR spectra of sugar alditols using a neural network. *J. Magnetic Reson.* 84, 212–217. doi: 10.1016/0022-2364(89)90021-8

Varnek, A., and Baskin, I. (2012). Machine learning methods for property prediction in chemoinformatics: quo vadis? *J. Chem. Inf. Model.* 52, 1413–1437. doi: 10.1021/ci200409x

Venkatasubramanian, V. (2019). The promise of artificial intelligence in chemical engineering: is it here, finally? *AIChE J.* 65, 466–478. doi: 10.1002/aic.16489

Wang, J., Olsson, S., Wehmeyer, C., Pérez, A., Charron, N. E., De Fabritiis, G., et al. (2019). Machine learning of coarse-grained molecular dynamics force fields. *ACS Central Sci.* 5, 755–767. doi: 10.1021/acscentsci.8b00913

Ward, L., and Wolverton, C. (2017). Atomistic calculations and materials informatics: a review. *Curr. Opin. Solid State Mater. Sci.* 21, 167–176. doi: 10.1016/j.cossms.2016.07.002

Wei, J. N., Duvenaud, D., and Aspuru-Guzik, A. (2016). Neural networks for the prediction of organic chemistry reactions. *ACS Central Sci.* 2, 725–732. doi: 10.1021/acscentsci.6b00219

Welborn, M., Cheng, L., and Miller, T. F. (2018). Transferability in machine learning for electronic structure via the molecular orbital basis. *J. Chem. Theory Comput.* 14, 4772–4779. doi: 10.1021/acs.jctc.8b00636

White, D., and Wilson, R. C. (2010). Generative models for chemical structures. *J. Chem. Inf. Model.* 50, 1257–1274. doi: 10.1021/ci9004089

Wu, Y., and Wang, G. (2018). Machine learning based toxicity prediction: from chemical structural description to transcriptome analysis. *Int. J. Mol. Sci.* 19:2358. doi: 10.3390/ijms19082358

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., et al. (2017). MoleculeNet: a benchmark for molecular machine learning. *arXiv e-prints*. Available online at: https://ui.adsabs.harvard.edu/abs/2017arXiv170300564W (accessed March 01, 2017).

Xia, R., and Kais, S. (2018). Quantum machine learning for electronic structure calculations. *Nat. Commun.* 9:4195. doi: 10.1038/s41467-018-06598-z

Xu, Y., Lin, K., Wang, S., Wang, L., Cai, C., Song, C., et al. (2019). Deep learning for molecular generation. *Future Med. Chem.* 11, 567–597. doi: 10.4155/fmc-2018-0358

Zaspel, P., Huang, B., Harbrecht, H., and Von Lilienfeld, O. A. (2019). Boosting quantum machine learning models with a multilevel combination technique: pople diagrams revisited. *J. Chem. Theory Comput.* 15, 1546–1559. doi: 10.1021/acs.jctc.8b00832

Zhang, P., Shen, L., and Yang, W. (2019). Solvation free energy calculations with quantum mechanics/molecular mechanics and machine learning models. *J. Phys. Chem. B* 123, 901–908. doi: 10.1021/acs.jpcb.8b11905

Zhou, Z., Kearnes, S., Li, L., Zare, R. N., and Riley, P. (2018). Optimization of molecules via deep reinforcement learning. *arXiv preprint arXiv*:1810.08678. doi: 10.1038/s41598-019-47148-x

Zhou, Z., Li, X., and Zare, R. N. (2017). Optimizing chemical reactions with deep reinforcement learning. *ACS Central Sci.* 3, 1337–1344. doi: 10.1021/acscentsci.7b00492

Zielinski, F., Maxwell, P. I., Fletcher, T. L., Davie, S. J., Di Pasquale, N., Cardamone, S., et al. (2017). Geometry optimization with machine trained topological atoms. *Sci. Rep.* 7:12817. doi: 10.1038/s41598-017-12600-3

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 28 January 2020
doi: 10.3389/fchem.2020.00016

A Combined Systematic-Stochastic Algorithm for the Conformational Search in Flexible Acyclic Molecules

David Ferro-Costas and Antonio Fernández-Ramos**

Center for Research in Biological Chemistry and Molecular Materials (CIQUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain

We propose an algorithm that is a combination of systematic variation of the torsions and Monte Carlo (or stochastic) search. It starts with a trial geometry in internal coordinates and with a set of preconditioned torsional angles, i.e., torsional angles at which minima are expected according to the chemical knowledge. Firstly, the optimization of those preconditioned geometries is carried out at a low electronic structure level, generating an initial set of conformers. Secondly, random points in the torsional space are generated outside the “area of influence” of the previously optimized minima (i.e., outside a hypercube about each minima). These random points are used to build the trial structure, which is optimized by an electronic structure software. The optimized structure may correspond to a new conformer (which would be stored) or to an already existing one. Initial torsional angles (and also final ones if a new conformer is found) are stored to prevent visiting the same region of the torsional space twice. The stochastic search can be repeated as many times as desired. Finally, the low-level geometries are recovered and used as the starting point for the high-level optimizations. The algorithm has been employed in the calculation of multi-structural quasi harmonic and multi-structural torsional anharmonic partition functions for a series of alcohols ranging from n-propanol to n-heptanol. It was also tested for the amino acid L-serine.

Keywords: conformational search, hindered rotors, torsional anharmonicity, stochastic methods, geometrical optimization

OPEN ACCESS

Edited by:
Jorge M. C. Marques,
University of Coimbra, Portugal

Reviewed by:
José Pedro Cerón-Carrasco,
Catholic University San Antonio of Murcia, Spain
Ekaterina Pas,
Monash University, Australia

****Correspondence:***
David Ferro-Costas
david.ferro@usc.es
Antonio Fernández-Ramos
qf.ramos@usc.es

Specialty section:
This article was submitted to Theoretical and Computational Chemistry,
a section of the journal Frontiers in Chemistry

Received: *18 October 2019*
Accepted: *08 January 2020*
Published: *28 January 2020*

Citation:
Ferro-Costas D and Fernández-Ramos A (2020) A Combined Systematic-Stochastic Algorithm for the Conformational Search in Flexible Acyclic Molecules. Front. Chem. 8:16. doi: 10.3389/fchem.2020.00016

1. INTRODUCTION

Flexible molecules have many conformational minima which can be easily reached by torsional motions of the molecular framework in the potential energy surface (PES). For the last few years, there are methods, as the multi-structural harmonic-oscillator (MS-HO) approximation (Zheng et al., 2011a), which take into account the characteristics of all these equilibrium structures. Specifically, the MS-HO method incorporates the rotational and vibrational (rovibrational) partition function of each of the conformers within the rigid-rotor harmonic-oscillator approximation. This is a substantial improvement over the one-well harmonic oscillator (1W-HO) approximation in which the structure of the absolute minimum is the only one to be considered (Ferro-Costas et al., 2018b).

Locating all conformers is just the first step toward the evaluation of more accurate rovibrational partition functions. For instance, it has been shown that MS-HO partition functions improve over 1W-HO ones (Ferro-Costas et al., 2018b), additionally torsional anharmonicity should be also included (Yu et al., 2011; Zheng et al., 2011b; Zheng and Truhlar, 2013) to increase the accuracy of the results. The most reliable methods that incorporate torsional anharmonicity can only be applied to a reduced number of torsional degrees of freedom (Fernández-Ramos, 2013) and they require more information of the PES than just the minima. For instance, the extended two-dimensional torsional method (E2DT) (Simón-Carballido et al., 2017), implemented in the Q2DTor program (Ferro-Costas et al., 2018a), needs a fine grid of points for the construction of the torsional PES. The procedure also includes the location of all stationary points (i.e., minima, saddle points and maxima in the 2D-PES).

Therefore, the amount of information needed from the PES depends on the method, and it is crucial to devise algorithms that allow an efficient construction of such PES. For example, when the number of torsional degrees of freedom is only 2, so the E2DT method can be applied, geometry scans at a regular number of points along the PES can be carried out. These scans involve partial optimizations in which all degrees of freedom are optimized except the two torsional modes. When the torsional global PES is calculated by systematic mapping, if possible, it is essential to reduce the number of points to be calculated. This reduction depends on molecular geometry aspects as conformational enantiomerism, internal symmetry of the rotors and molecular symmetry. The rules to replicate points of a PES under some symmetry conditions are given in Ferro-Costas et al. (2018a). As the number of torsional degrees of freedom increases, the amount of information needed from the PES should be reduced in order to keep the problem tractable. For those cases, the multi-structural torsional method is a good choice (Zheng et al., 2012, 2013), because the model is built assuming that the only information at hand is the set of conformational minima.

This work is concerned with the search of conformational structures in the torsional PES of flexible acyclic molecules with more than 2 torsions (typically up to 10). Having the equilibrium geometries, it is possible to calculate accurate rovibrational partition functions in a wide range of temperatures. In this sense, the algorithm is not limited to the search of the most stable equilibrium structures, O'Boyle et al. (2011) which are the only ones that are relevant at low temperatures. It seeks for *all* conformers, because they are required for the calculation of partition functions at high temperatures and for the evaluation of torsional anharmonicity. Unfortunately, this algorithm cannot deal with large biological systems or with conformations originated from ring puckering (Kolossváry and Guida, 1996; Watts et al., 2010). For that purpose there is an extense list of algorithms and programs (see Loferer et al., 2007; Friedrich et al., 2019 and references therein).

The algorithm here presented is a combination of a systematic method that locates intuitively expected conformers plus a Monte Carlo method that finds unanticipated ones. A detailed description of the algorithm is given in the following section. The series of alcohols ranging from n-propanol to n-heptanol and the amino acid L-serine have been selected to test the algorithm.

2. DESCRIPTION OF THE ALGORITHM

The target systems for this algorithm are flexible acyclic molecules characterized by t dihedral angles. Internal rotations about these dihedrals guide the system toward different conformations; each of them being represented by a t-dimensional point $\Phi = (\phi_1, \cdots, \phi_\tau, \cdots, \phi_t)$, where the τ-th dihedral angle runs from 1 to t.

The various geometries involved in the algorithm are:

- Φ^{R}: the reference geometry, i.e., the initial geometry provided by the user.
- Φ^{G_1}: a guess geometry during the systematic search. The total number of structures generated is K_1 of which $K_1^\star$ are the ones that pass the tests (see section 2.1) and turn into trial geometries. Notice that $K_1^\star \leq K_1$.
- Φ^{G_2}: a guess geometry during the stochastic search. The total number of geometries generated is K_2 of which $K_2^\star$ are the ones that pass the tests (see section 2.1) and turn into trial geometries. Notice that $K_2^\star \leq K_2$.
- $\Phi^0_{k^\star}$: the $k^\star$-th trial geometry, $k^\star = 1, \ldots, K^\star$; $K^\star = K_1^\star + K_2^\star$. The pool of trial geometries is represented by $\{\Phi^0_{k^\star}\}$.
- $\Phi^\star$: a trial geometry to be optimized.
- $\Phi^{\star,\mathrm{opt}}$: a trial geometry $\Phi^\star$ after optimization.
- Φ^{eq}_j: : the j-th equilibrium conformer, $j = 1, \ldots, J$. The pool of such conformers is represented by $\{\Phi^{\mathrm{eq}}_j\}$.
- Φ^{st}_p: the p-th stored point, $p = 1, \ldots, P$; $P = K_1^\star + K_2^\star + J$. The pool of stored points is the union of the previous two sets, $\{\Phi^{\mathrm{st}}_p\} = \{\Phi^{\mathrm{eq}}_j\} \cup \{\Phi^0_{k^\star}\}$.

The setup of the algorithm is schematically shown in the flux diagram of **Figure 1**. It starts with a reference geometry given in the Z-matrix format where the t target torsions must be defined unambiguously. Only in this manner, it is possible to define the Φ^{R} torsional point univocally. Otherwise, the torsional analysis cannot be carried out. The algorithm consists of two well differentiated searching methods: systematic and stochastic.

2.1. The Tests

A guess structure, either Φ^{G_1} or Φ^{G_2}, in general denoted as Φ^{G}, must complete two tests before being considered a trial geometry $\Phi^\star$:

1. *The connectivity test*: excludes structures having unphysical bond lengths (for instance, structures with superimposed atoms) or structures with different connectivity than the reference geometry. The detection and exclusion of these type of structures is carried out through the adjacency (or connectivity) matrix of the guess geometry, $\mathbf{A}(\Phi^{\mathrm{G}})$, which is compared to that of the reference structure, $\mathbf{A}(\Phi^{\mathrm{R}})$. Only if these two matrices are equal:

$$\mathbf{A}(\Phi^{\mathrm{R}}) = \mathbf{A}(\Phi^{\mathrm{G}}) \tag{1}$$

the guess geometry passes the test. We highlight the importance of generating an adequate reference structure, as its connectivity matrix is used to accept or discard guess geometries.

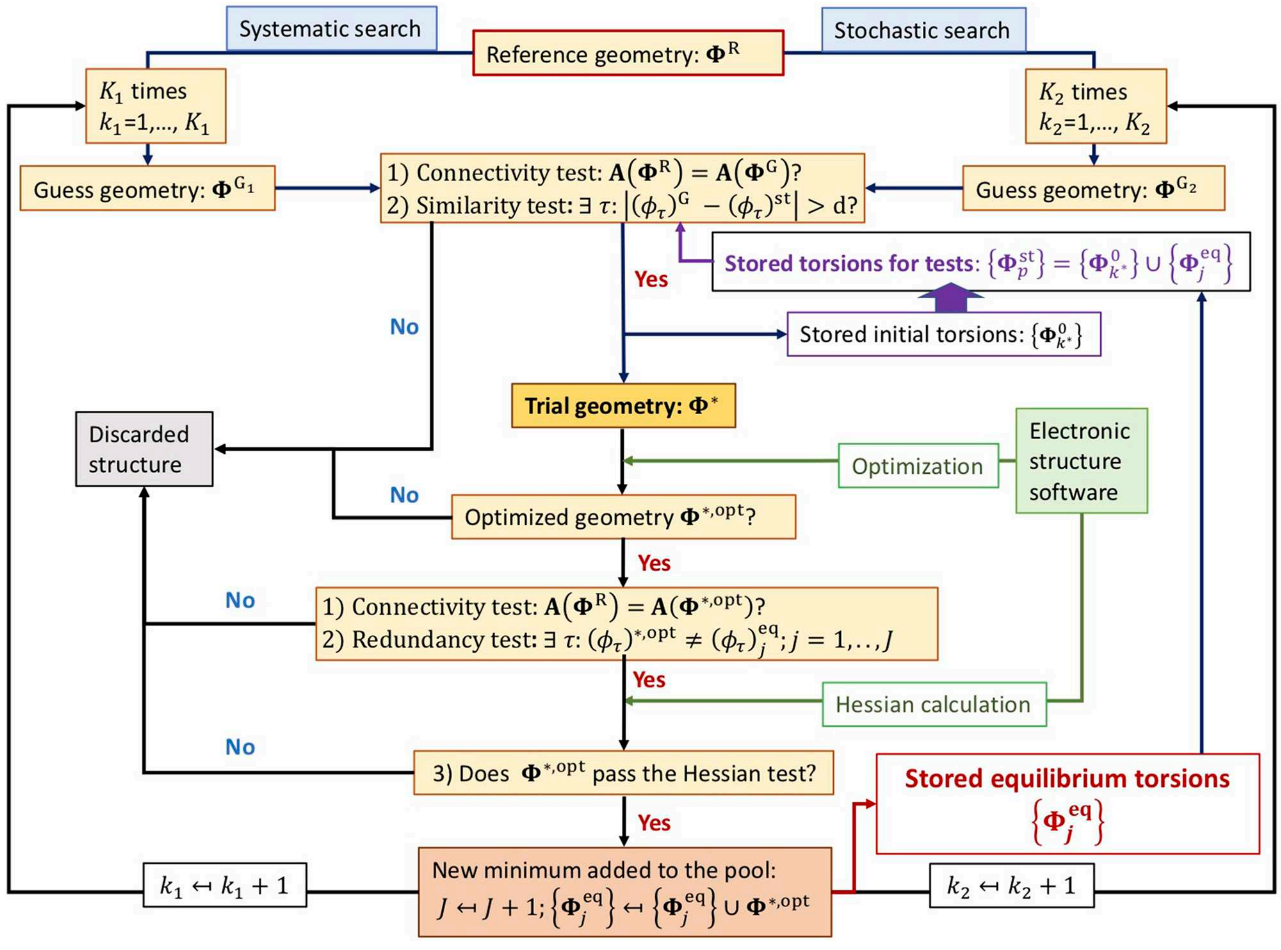

Systematic search
Reference geometry: $\mathbf{\Phi}^{\mathrm{R}}$
Stochastic search
K_1 times k_1=1,..., K_1
K_2 times k_2=1,..., K_2
Guess geometry: $\mathbf{\Phi}^{\mathrm{G}_1}$
1) Connectivity test: $\mathbf{A}(\mathbf{\Phi}^{\mathrm{R}}) = \mathbf{A}(\mathbf{\Phi}^{\mathrm{G}})$?
2) Similarity test: $\exists\,\tau$: $|(\phi_\tau)^{\mathrm{G}} - (\phi_\tau)^{\mathrm{st}}| > \mathrm{d}$?
Guess geometry: $\mathbf{\Phi}^{\mathrm{G}_2}$
Stored torsions for tests: $\{\mathbf{\Phi}_p^{\mathrm{st}}\} = \{\mathbf{\Phi}_{k^*}^{0}\} \cup \{\mathbf{\Phi}_j^{\mathrm{eq}}\}$
Yes
No
Stored initial torsions: $\{\mathbf{\Phi}_{k^*}^{0}\}$
Trial geometry: $\mathbf{\Phi}^*$
Discarded structure
Optimization
Electronic structure software
No
Optimized geometry $\mathbf{\Phi}^{*,\mathrm{opt}}$?
Yes
No
1) Connectivity test: $\mathbf{A}(\mathbf{\Phi}^{\mathrm{R}}) = \mathbf{A}(\mathbf{\Phi}^{*,\mathrm{opt}})$?
2) Redundancy test: $\exists\,\tau$: $(\phi_\tau)^{*,\mathrm{opt}} \neq (\phi_\tau)_j^{\mathrm{eq}}; j = 1,..,J$
Yes
Hessian calculation
No
3) Does $\mathbf{\Phi}^{*,\mathrm{opt}}$ pass the Hessian test?
Stored equilibrium torsions $\{\mathbf{\Phi}_j^{\mathrm{eq}}\}$
Yes
$k_1 \leftarrow k_1 + 1$
New minimum added to the pool:
$J \leftarrow J + 1; \{\mathbf{\Phi}_j^{\mathrm{eq}}\} \leftarrow \{\mathbf{\Phi}_j^{\mathrm{eq}}\} \cup \mathbf{\Phi}^{*,\mathrm{opt}}$
$k_2 \leftarrow k_2 + 1$

FIGURE 1 | Flux diagram of the conformer search algorithm.

2. *The similarity test*: performs a comparison between the guess point and the pool of the P stored points. If Φ^{G} falls outside of all the hypercubes generated about each stored point, the geometry is accepted. For hypercubes with edge size of $2d$, this test is positive if:

$$\exists\,\tau : |(\phi_\tau)^{\mathrm{G}} - (\phi_\tau)_p^{\mathrm{st}}| > d, \quad p = 1, \cdots, P \tag{2}$$

An optimized trial geometry, $\Phi^{\star,\mathrm{opt}}$, must complete three tests.

1. *The connectivity test*: assures that the optimized and reference geometries share the same connectivity.
2. *The redundancy test*: compares the current geometry with the pool of the J optimized equilibrium torsions. If

$$\exists\,\tau : (\phi_\tau)^{\star,\mathrm{opt}} \neq (\phi_\tau)_j^{\mathrm{eq}}, \quad j = 1, \cdots, J \tag{3}$$

   then $\Phi^{\star,\mathrm{opt}}$ is a candidate for being a new equilibrium structure in the torsional PES.
3. *The Hessian test*: assures that the optimized geometry is a new minimum. The electronic structure software is used to calculate the Hessian matrix; if the normal-mode frequencies of the diagonalized matrix are all real, the optimized structure is a minimum.

Notice that the order in which the tests are carried out is important because if the optimized structure fails to pass the first two tests, no time is lost in the evaluation of the Hessian matrix.

2.2. Systematic Search

The first part of the algorithm consists of a systematic search that makes use of a pool of K_1 initial structures, each of them characterized by a set of torsional angles Φ^{G_1} which have their origin on basic molecular structure analysis. If there are P_τ initial chemical-intuitive guesses for a given torsion τ, the total number of preconditioned guesses is:

$$K_1 = \prod_{\tau=1}^{t} P_\tau \tag{4}$$

For instance, for a four sp^3 carbon linear chain, the expected location of the minima is at dihedral angles of 180° and ±60°, which correspond to the *anti* (antiperiplanar ot T) and *gauche* (synclinal or G±) positions, and therefore $P_\tau = 3$ for the torsion. Notice that only dihedral angles that generate new distinguishable structures should be included. In this context, methyl groups should be ignored because its internal rotation only generates indistinguishable structures. For instance, in the case of n-butanol we only need to consider three torsions ($t = 3$), each of them with three intuitive positions (T, G+, and G−), that is $P_1 = P_2 = P_3 = 3$. Therefore, the number of geometries to be generated within the pool is $K_1 = 27$.

During the generation of the initial structures, the algorithm should take into account the characteristics of the molecular geometry, as for instance, molecular symmetry. Returning to the previous example, the molecule of n-butanol has one structure given by the dihedrals (TTT) which has a plane of symmetry and, therefore, it belongs to the C_s point group symmetry. As a consequence, all structures, with exception of (TTT), have conformational enantiomers, that is, distinguishable optical isomers with the same electronic structure properties. Therefore, it is sufficient to locate one of the two isomers. The conformational enantiomer of the Φ structure is the $-\Phi$ structure (i.e., the value of the dihedral angle for each torsion is set to $360^\circ - \phi_\tau$). For instance, structure (TG+G-) has structure (TG-G+) as enantiomer. Consequently, only 14 of the 27 initial structures need to be tried. In general, for a molecule with a plane of symmetry, the initial number of structures of the preconditioned systematic search is reduced to $(K_1 + 1)/2$. The rest of the structures are automatically generated from the calculated ones.

Each of the K_1 structures leads to a guess point, Φ^{G_1}, which is the current candidate to turn into a new minimum in the PES upon geometry optimization. If this point fails to pass the *connectivity* or the *similarity* tests, it is discarded. Otherwise, Φ^{G_1} is a suitable structure to be optimized by the electronic structure software:

$$\Phi^\star \leftarrow \Phi^{G_1} \tag{5}$$

and the $\{\Phi^0_{k^\star}\}$ pool is updated:

$$\{\Phi^0_{k^\star}\} \leftarrow \{\Phi^0_{k^\star}\} \cup \{\Phi^\star\} \tag{6}$$

During the systematic search the *similarity test* only involves $\{\Phi^{\rm eq}_j\}$ and not $\{\Phi^{\rm st}_p\}$. The reason is that the hypercubes generated from these structures do not overlap.

If $\Phi^\star$ successfully converges to an equilibrium structure, $\Phi^{\star,{\rm opt}}$, and passes all the required tests, then the resulting geometry is added to the list of conformers and the pool of equilibrium torsions is updated:

$$\{\Phi^{\rm eq}_j\} \leftarrow \{\Phi^{\rm eq}_j\} \cup \{\Phi^{\star,{\rm opt}}\} \tag{7}$$

2.3. Stochastic Search

The algorithm can perform a series of K_2 cycles performing a Monte Carlo search after the systematic procedure. At every cycle, the algorithm generates t random numbers, which are the components of the torsional point Φ^{G_2}. Once they are generated, the procedure follows the same pattern as the systematic search. In this case, the *similarity test* accesses to the $\{\Phi^{\rm st}_p\}$ pool, not just to the equilibrium structures, because it cannot be assured that the new point falls outside of the "area of influence" of previous trial geometries.

We highlight that it is possible to run several batches of the Monte Carlo search, each of them with different specifications for the hypercube edge size ($2d$).

2.4. Electronic Structure Calculations

The algorithm assumes that the initial set of conformers will be obtained at a low electronic structure level (LL). Those equilibrium geometries can be used as the starting point of high-level (HL) electronic structure calculations. Therefore, the LL calculations should produce a torsional PES which, at least qualitatively, has a similar topology than the HL torsional PES. For molecules of the size presented in this work, Hartree-Fock (HF) calculations are affordable as the LL. Other reason to choose HF as the LL is that they tend to overestimate torsional barriers, as well as to produce more minima than electronic correlated methods. However, we notice that the LL search is not restricted to HF. Molecular mechanics, semiempirical methods or other *ab initio* methods could be used as LL, depending on the molecular system and on the available computational resources. In principle, the algorithm was designed for locating minima in the torsional PES of flexible acyclic systems with up to 10 torsions. Bigger systems would require substantial computational cost. Two straightforward ways of reducing computer time are the parallelization of the algorithm and/or the use of inexpensive LL methods.

For the n-alcohols series and L-serine, HF/3-21G was chosen as the LL method. For n-alcohols the set of the $\{\Phi^{\rm eq}_j\}$ conformers was re-optimized employing the MPWB1K functional (Zhao et al., 2004) in combination with the 6-31+G(d,p) basis set (Hehre et al., 1972). In the case of serine, B3LYP/6-31++G** was the method of choice with the objective of establishing a direct comparison with previous calculations (Najbauer et al., 2015). Geometry optimizations and frequency calculations were carried out with the *Gaussian 09* package (Frisch et al., 2004).

3. MULTI-STRUCTURAL PARTITION FUNCTIONS

It has been shown that for flexible molecules the incorporation of multiple conformers may have a substantial impact in the magnitude of the partition functions, thermochemical properties and thermal rate constants. The algorithm has been designed to obtain *all* the torsional PES minima of the molecule. After this information is accessible, it is possible to calculate multi-structural (MS) partition functions, i.e., partition functions that include multiple torsional conformers. Here we are concerned with the calculation of the MS partition functions using the following approximations: the harmonic-oscillator (MS-HO), the quasi-harmonic (MS-QH), and the coupled torsional anharmonic (MS-T(C); Zheng and Truhlar, 2013). Any of these methods, named in increasing order of accuracy, provides more reliable values of thermochemical properties than methods based on just one well. In the MS-HO approximation the rovibrational partition function is given by

$$Q^{\mathrm{MS-HO}} = \sum_{j_c}^{J_c} Q_{j_c}^{\mathrm{rot}} Q_{j_c}^{\mathrm{HO}} e^{-\beta U_{j_c}} \tag{8}$$

where J_c is the total number of conformers and U_{j_c} is the relative energy of conformer j_c relative to the global minimum. Without lost of generality, it is possible to sum just over *all* the distinguishable structures J that are not conformational enantiomers

$$Q^{\mathrm{MS-HO}} = \sum_{j}^{J} w_j Q_j^{\mathrm{rot}} Q_j^{\mathrm{HO}} e^{-\beta U_j} \tag{9}$$

where $w_j = 1$ if the j structure is unique and $w_j = 2$ if it has a conformational enantiomer. The rigid rotor rotational partition function Q_j^{rot} is given by

$$Q_j^{\mathrm{rot}} = \frac{8\pi^2}{\sigma_{\mathrm{rot},j}} \left(\frac{1}{2\pi\hbar^2\beta} \right)^{3/2} \sqrt{I_{1,j}^{\mathrm{rot}} I_{2,j}^{\mathrm{rot}} I_{3,j}^{\mathrm{rot}}} \tag{10}$$

where $\hbar$ is the Planck's constant divided by 2π, and $\beta = (k_{\mathrm{B}}T)^{-1}$, with k_{B} being the Boltmann's constant and T the temperature; $\sigma_{\mathrm{rot},j}$ is the symmetry number of rotation (Fernández-Ramos et al., 2007) and $I_{i,j}^{\mathrm{rot}}$ ($i = 1$, 2 or 3) is the i-th principal moment of inertia of conformer j.

The harmonic oscillator partition Q_j^{HO} is

$$Q_j^{\mathrm{HO}} = \tilde{Q}_j^{\mathrm{HO}} e^{-\beta \mathcal{E}_j^{\mathrm{HO}}} \tag{11}$$

where

$$\tilde{Q}_j^{\mathrm{HO}} = \prod_{m=1}^{3N-6} \frac{1}{1 - e^{-\beta\hbar\omega_{m,j}}} \tag{12}$$

is the HO vibrational partition function calculated by taking the zero-point energy (ZPE) as the reference energy, which is given by

$$\mathcal{E}_j^{\mathrm{HO}} = \sum_{m=1}^{3N-6} \frac{1}{2}\hbar\omega_{m,j} \tag{13}$$

where N is the number of atoms, and $\omega_{m,j}$ is the HO frequency of the m-th normal mode in the j-th conformer. A variant of the MS-HO partition function is the MS-QH one, in which the harmonic frequencies are multiplied by a scale parameter λ^{ZPE} which is dependent on the electronic structure method and that it was previously parametrized to reproduce experimental ZPEs. Thus,

$$Q_j^{\mathrm{QH}} = \tilde{Q}_j^{\mathrm{QH}} e^{-\beta \mathcal{E}_j^{\mathrm{QH}}} \tag{14}$$

$$\tilde{Q}_j^{\mathrm{QH}} = \prod_{m=1}^{3N-6} \frac{1}{1 - e^{-\beta\hbar\lambda^{\mathrm{ZPE}}\omega_{m,j}}} \tag{15}$$

$$\mathcal{E}_j^{\mathrm{QH}} = \lambda^{\mathrm{ZPE}} \sum_{m=1}^{3N-6} \frac{1}{2}\hbar\omega_{m,j} \tag{16}$$

and therefore

$$Q^{\mathrm{MS-QH}} = \sum_{j}^{J} w_j Q_j^{\mathrm{rot}} Q_j^{\mathrm{QH}} e^{-\beta U_j} \tag{17}$$

The MS-T(C) rovibrational partition function includes torsional anharmonicity on the HO or QH partition functions through a multiplicative factor $F_{\mathrm{cl},j}^{\mathrm{MS-T(C)}}$. For the QH case, it is given by

$$Q^{\mathrm{MS-T(C)}} = \sum_{j=1}^{J} Q_j^{\mathrm{rot}} Q_j^{\mathrm{QH}} F_{\mathrm{cl},j}^{\mathrm{MS-T(C)}} \tag{18}$$

where

$$F_{\mathrm{cl},j}^{\mathrm{MS-T(C)}} = \prod_{\eta=1}^{t} f_{j,\eta} = \prod_{\eta=1}^{t} \frac{q_{j,\eta}^{\mathrm{RC(C)}}}{q_{j,\eta}^{\mathrm{CHO(C)}}} \tag{19}$$

The $f_{j,\eta}$ factors are expressed as the ratio between the classical reference anharmonic (RC) and classical harmonic oscillator (CHO) torsional partition functions. Although the reference classical partition function involves some approximations, it incorporates couplings in the kinetic and potential energies between the torsions. Therefore, in flexible systems with multiple torsional modes, the MS-T(C) entails a substantial improvement over the MS-QH method.

4. RESULTS AND DISCUSSION

The automatic protocol presented in this work was adopted to study the n-alcohols from 3 (n-propanol) to 7 (n-heptanol) carbon atoms. For the calculation of the partition functions, the frequencies were scaled by the recommended factor $\lambda^{\mathrm{ZPE}} = 0.951$ (Alecu et al., 2010). With the exception of n-propanol and n-butanol, the number of previous studies on the conformations of n-alcohols is scarce. Thus, additionally to the seek for conformational minima of n-alcohols, we have benchmarked our algorithm against a previous study on the conformations of the amino acid L-serine, a molecule presenting several functional groups (section 4.2). However, we will center our attention on the n-alcohols when discussing about the efficiency of the algorithm.

4.1. Efficiency of the Algorithm

A summary about the efficiency of the algorithm for n-alcohols is provided in **Table 1**. In it, J_1 and J_2 represent the number of conformers found in the systematic and in the Monte-Carlo searchings, respectively. Similarly, J_{LL} and J_{HL} are the total number of conformers found at LL and HL, respectively. We highlight that the number of conformers shown in **Table 1** excludes enantiomeric structures. Consequently, the total number of conformers is given by $2J_{\mathrm{HL}} - 1$. The relative total energy of the HL conformers is represented in **Figure 2** for the five n-alcohols. We refer to the **Supplementary Material** for a list containing the electronic energy and Cartesian coordinates of all the HL conformers.

For the case of n-heptanol, according to the chemical intuition and excluding enantiomers, a total of $(3^6 + 1)/2 = 365$

TABLE 1 | Number of conformers obtained for the n-alcohol series that illustrate the efficiency of the algorithm.

Alcohol	t	K_1	$K_1^\star$	J_1	K_2	$K_2^\star$	J_2	J_{LL}	J_{HL}
n-propanol	2	5	5	5	200	80	0	5	5
n-butanol	3	14	14	14	200	167	1	15	15
n-pentanol	4	41	39	38	400	377	15	53	48
n-hexanol	5	122	110	106	4,500	3,988	59	165	153
n-heptanol	6	365	307	297	5,500	4,664	192	489	465

See text for details about nomenclature.

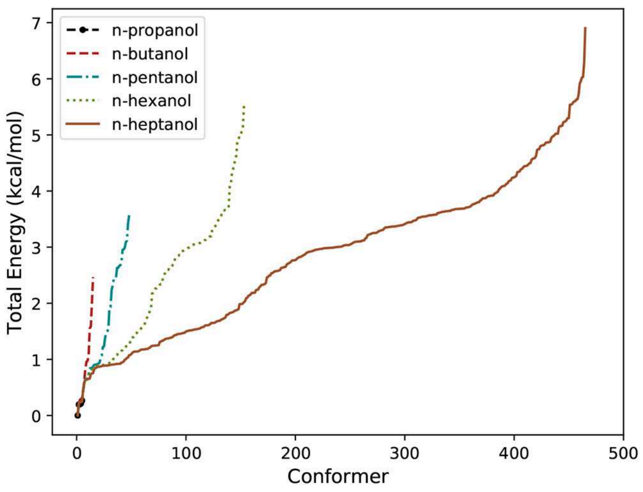


FIGURE 2 | Number of conformers of n-alcohols and their relative total energy.

conformers are expected. The algorithm discarded 58 of them as a result of very strained geometries which did not pass the *connectivity test*. Therefore, geometric optimizations were performed on 84% of the initial geometries of which 97% of them led to a new conformer. From these results, we can state that the systematic search is very efficient because almost every geometry optimization that was carried out translated into a new conformer. This result is not surprising as the starting geometries of the systematic search arise from well-established chemical knowledge.

The performance of the stochastic search is more difficult to estimate. About 15% of the generated geometries were immediately discarded through the *connectivity test* saving a considerable amount of computational time. Notice that "all" minima with torsional angles lying close to the preconditioned values have already been found, so only about 4% of the geometry optimizations led to new conformers. However, this result should not be taken as poor performance of the algorithm, but as an inherent difficulty associated with the search of new conformers in partially explored PESs. Every new batch of calculations produces less new minima, and the algorithm is stopped when no new conformers are found. However, even in this situation, the location of all the conformers is not guaranteed.

Regarding to the HL optimizations from the LL geometries, we observe that the procedure is quite effective: for the five n-alcohols, more than 90% of the LL geometries leaded to a new HL conformer.

4.2. Benchmarking

Studies regarding to the conformational flexibility of n-alcohols beyond n-butanol are scarce in the literature. Even for n-butanol, some of these previous studies (see Black and Simmie, 2010) pointed toward the existence of 14 conformers (27 considering enantiomers), which are the number of conformers encountered after a systematic search. However, the total number of conformers is 15. This last conformer appears after a stochastic search.

Chen at al. (2015) claimed that n-pentanol has 41 minima, which are the hypothetical number of conformers generated by T, G+ and G- configurations for each of the torsions. Our algorithm, discarded two of them in the systematic search and encountered 38 conformers at HL. The stochastic search located another 10 conformers to reach a total of 48 conformers. The algorithm may discard some initial geometries if they do not pass the connectivity test; however, if there are minima close to these strained geometries, they will be encountered during the Monte Carlo search.

For n-hexanol there is a very recent work by Vaskivskyi et al. (2019), who made a systematic search starting from 122 structures and found 111 different conformers. Our algorithm located a total of 153 minima at the HL, 106 in the systematic search and 47 in the Monte Carlo search.

To the best of our knowledge this is the first work dealing with the conformational flexibility of n-heptanol.

Najbauer et al. (2015) reported the 14 conformers of L-serine with the lowest Gibbs free energies at 0 K, $\Delta G^{\text{o}}_{0\text{K}}$, calculated at the B3LYP/6-31++G** level. Our algorithm found a total of 72 LL conformers, number that was reduced to 60 (listed in the **Supplementary Material**) after the HL reoptimizations, also at the B3LYP/6-31++G** level. Of the total number of conformers obtained at the HL, 32 of them are within the range of free energies reported by Najbauer et al. (2015) (see **Table 2**). Specifically, Najbauer *et al* missed 18 conformers within a free energy window of 4 kcal/mol.

4.3. Multiple Wells and Torsional Anharmonicity

The number of conformers increases with the size of the system, as shown in **Table 1**, although this does not imply that all of them are required in the calculation of thermodynamic properties. The importance of each of the j-th conformers can be estimated by its contribution, χ_j, to the MS-QH partition function:

$$\chi_j = \frac{w_j Q_j^{\text{rot}} Q_j^{\text{QH}} e^{U_j/k_B T}}{\sum_j w_j Q_j^{\text{rot}} Q_j^{\text{QH}} e^{U_j/k_B T}} = \frac{w_j e^{-G_j/k_B T}}{\sum_j w_j e^{-G_j/k_B T}} \tag{20}$$

where G_j is the rovibrational Gibbs free energy of the j-th conformer:

$$G_j = U_j - k_B T \ln \left[Q_j^{\text{rot}} Q_j^{\text{QH}} \right] \tag{21}$$

TABLE 2 | L-serine low-energy conformers sorted out according to their relative harmonic Gibbs free energies at 0 K (in kcal/mol) calculated at the B3LYP/6-31++G** level.

This work	$(\phi_1, \phi_2, \phi_3, \phi_4, \phi_5)$	$\Delta G^o_{0\ K}$	Najbauer et al.
1	(178, 289, 181, 044, 278)	0.00	1
2	(355, 145, 293, 082, 092)	0.17	2
3	(356, 145, 064, 305, 092)	0.56	3
4	(180, 306, 294, 314, 217)	0.92	4
5	(183, 100, 177, 043, 283)	1.56	5
6	(001, 108, 181, 179, 149)	1.61	6
7	(183, 194, 296, 068, 308)	1.63	7
8	(357, 145, 175, 185, 091)	1.68	8
9	(003, 106, 182, 281, 148)	1.71	9
10	(180, 181, 058, 291, 297)	2.12	10
11	(181, 306, 297, 187, 312)	2.20	
12	(358, 139, 176, 272, 097)	2.27	
13	(184, 323, 299, 083, 317)	2.39	11
14	(178, 161, 066, 295, 190)	2.47	
15	(179, 302, 072, 303, 294)	2.64	12
16	(179, 127, 290, 316, 206)	2.73	
17	(184, 214, 292, 062, 085)	2.74	
18	(181, 318, 075, 296, 208)	2.77	
19	(179, 126, 292, 181, 309)	3.06	13
20	(006, 316, 065, 189, 224)	3.12	
21	(177, 282, 182, 283, 039)	3.22	
22	(177, 280, 179, 174, 045)	3.25	
23	(358, 144, 170, 077, 090)	3.26	
24	(177, 043, 288, 066, 310)	3.27	
25	(184, 321, 174, 168, 192)	3.30	
26	(180, 169, 175, 172, 178)	3.32	
27	(177, 261, 051, 066, 287)	3.36	
28	(179, 245, 058, 190, 286)	3.48	
29	(177, 293, 074, 306, 041)	3.65	
30	(182, 125, 067, 309, 058)	3.72	
31	(007, 318, 061, 083, 221)	3.82	
32	(183, 318, 176, 278, 187)	3.97	14

*See the **Supplementary Material** for the labeling of the five dihedral angles, $(\phi_1, \phi_2, \phi_3, \phi_4, \phi_5)$. Last column labels the 14 lowest-energy conformers according to the work of Najbauer et al. (2015).*

TABLE 3 | Minimum number of HL conformer needed to achieve $\sum \chi_j \geq 0.9$ at 300, 1,000, and 2,500 K.

	300 K		1,000 K		2,500 K	
System	J_1	J_2	J_1	J_2	J_1	J_2
n-propanol	5	0	5	0	5	0
n-butanol	10	0	12	0	13	0
n-pentanol	20	0	29	1	31	5
n-hexanol	40	0	75	12	84	20
n-heptanol	94	0	199	45	218	81

The minimum number is split into conformers obtained by the systematic (J_1) and stochastic (J_2) algorithms, respectively.

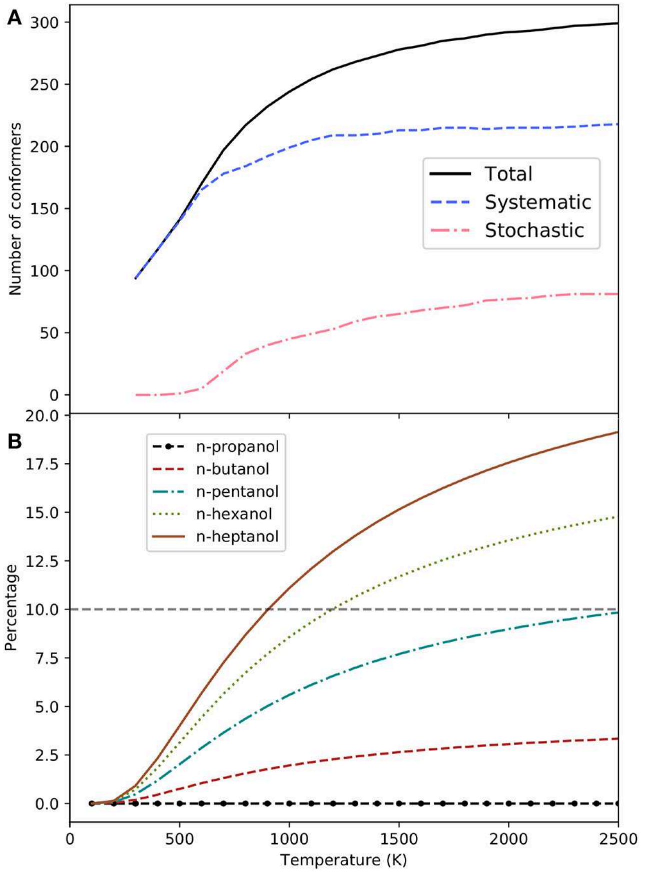


FIGURE 3 | (A) Minimum number of n-heptanol conformers needed to achieve $\sum_j \chi_j \geq 0.9$ (solid line). Conformers collected by the systematic (dashed line) and stochastic (dash-dotted line) method are also indicated. **(B)** Contribution to the MS-QH partition function of the conformers found during the stochastic search of n-alcohols.

We highlight that χ_j also represents the relative population of the j-th conformer.

The conformers of each alcohol can be sorted out according to their χ_j value in such a way that $\chi_j \geq \chi_{j+1}$. Considering an error of 10% as acceptable in the evaluation of the MS-QH partition functions, it is possible to estimate the minimum number of conformers needed to recover 90% of the partition function. This number can be factorized into conformers obtained by the systematic method and into conformers obtained by the stochastic algorithm. The analysis has been performed in the range of temperatures between 100 and 2,500 K and it can be found in **Table 3**. For the specific case of n-heptanol the values are plotted in **Figure 3A**.

At the low temperatures regime, the number of conformers that contribute to the free energy is small, and most of them belong to the pool of conformations obtained in a systematic manner. In fact, the stochastic method is not needed for n-propanol and n-butanol in the whole range of temperatures studied here. For n-heptanol (**Figure 3A**), we notice that (i) the importance of conformers obtained by the stochastic method is negligible at temperatures smaller than 700 K, (ii) even at higher temperatures, only 64% of the total number of conformers are needed to recover 90% of the MS-QH partition function.

It is obvious that the stochastic algorithm is less efficient than the systematic procedure. Consequently, locating conformers arising from the stochastic search requires higher computational cost. Unfortunately, this search is compulsory for the largest alcohols studied. The equilibrium structures retrieved by the stochastic search account for 0, 7, 21, 31, and 37% of the total for n-propanol to n-heptanol HL structures, respectively. As expected (see **Figure 3B**), their contribution increases with temperature, as well as with system size. However, if we concede deviations up to 10% in the partition function, these conformations are essential for both n-hexanol (from 1,100 K) and n-heptanol (from 900 K), but not for the small n-alcohols.

In order to study the repercussion of the multiple wells and torsional anharmonicity in the n-alcohol series, we have employed the MsTor program (Zheng et al., 2013), which can handle the calculation of MS-QH and MS-T(C) and partition functions. The effect of multiple wells was analyzed through the $Q^{\mathrm{MS-QH}}$ and $Q^{\mathrm{1W-QH}}$ ratio, where 1W-QH refers to the quasi-harmonic version of the absolute minimum. The evolution of this ratio with temperature is plotted in **Figure 4A**. The chart shows that the one-well approximation is unsatisfactory even at very low temperatures. The impact of the system size is also substantial; the single conformer approximation turned out worst with longer carbon chains. For instance, at 1,000 K, this ratio increases from 8 to 164 when moving from n-propanol to n-heptanol.

We have also analyzed the variation of the $Q^{\mathrm{MS-T(C)}}/Q^{\mathrm{MS-QH}}$ ratio with temperature (**Figure 4B**). Both partition functions are multi-structural, so they include the whole set of conformers. Therefore, the ratio shows the impact of the torsional anharmonicity in the partition functions. Torsional anharmonicity is slightly smaller than the unity at low temperatures (between 0.8 and 1.0) and increases to about 1.4 for n-hexanol and n-heptanol at 700 K. At higher temperatures the ratio declines again. The reason for this behavior is that at high temperatures the density of states of the hindered rotor partition function diminish with increasing temperature, whereas the density of states of the harmonic oscillator remains constant. Therefore, it is crucial to incorporate torsional anharmonicity in the harmonic partition function to retrieve the correct high temperature behavior (**Figure 4C**). As a general rule, there are two factors that require careful consideration: number of conformations, and torsional anharmonicity. We subscribe to the comment of S. J. Klippenstein, who in a recent review stated (Klippenstein, 2017) "*Historically, the uncertainties in theoretical predictions have been dominated by uncertainties in the barrier height predictions, but this is no longer the case. Uncertainties in the partition function evaluations are now often of comparable or even larger magnitude*."

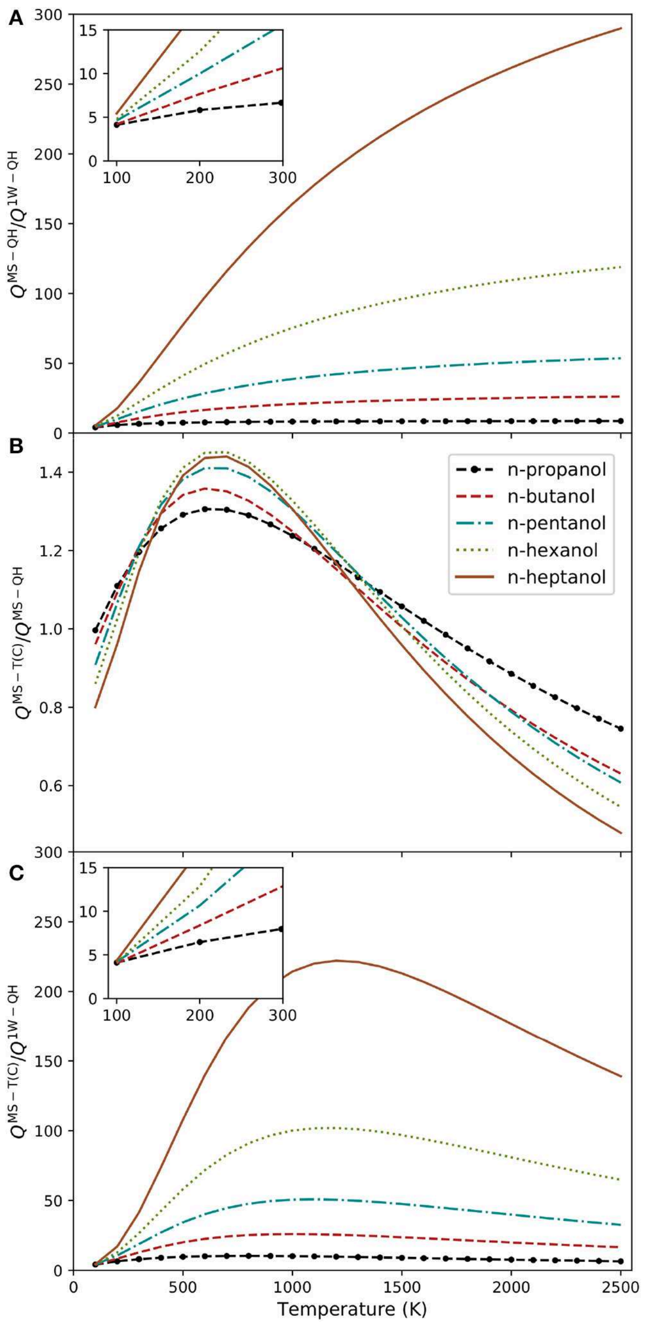


FIGURE 4 | Alcohols partition function ratios plotted at several temperatures: **(A)** $Q^{\mathrm{MS-QH}}/Q^{\mathrm{1W-QH}}$; **(B)** $Q^{\mathrm{MS-T(C)}}/Q^{\mathrm{MS-QH}}$; and **(C)** $Q^{\mathrm{MS-T(C)}}/Q^{\mathrm{1W-QH}}$.

5. CONCLUSIONS

In this work we have presented a combined algorithm able to locate *all* torsional conformers of medium-size acyclic molecules. The algorithm accepts two different strategies for the generation of trial structures: a systematic one, based on the chemical knowledge, and a stochastic one. The torsional PES is efficiently visited, avoiding previously explored areas.

This algorithm was tested in the series of n-alcohols ranging from n-propanol to n-heptanol, as well as in L-serine. We have encountered that the number of conformers arising from the stochastic search is not negligible for n-hexanol and n-heptanol. At the low temperatures regime the contribution to the partition function of the conformers found during the stochastic search is negligible. However, at medium/high temperatures, their exclusion can lead to significant errors. In combination with the MSTor program, the algorithm allows an efficient computation of the MS-QH and MS-T(C) partition functions. The results indicate that the one-well approximation substantially underestimates the magnitude of the partition function when compared with the multi-structural methods. In the case of L-serine, the algorithm was able to locate additional conformers to those described in recent works. In fact, within a range of 4 kcal/mol, the algorithm was able to locate 32 conformers, unlike to the 14 conformers previously reported.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

DF-C and AF-R have equally contributed to the development of the algorithm, the DFT calculations and the evaluation of multi-structural partition functions for the n-alcohols and L-serine here presented.

FUNDING

Financial support from the Consellería de Cultura, Educación e Ordenación Universitaria (Axuda para Consolidación e Estructuración de unidades de investigación competitivas do Sistema Universitario de Galicia, Xunta de Galicia ED431C 2017/17 & Centro singular de investigación de Galicia acreditación 2016-2019, ED431G/09) and the European Regional Development Fund (ERDF) is gratefully acknowledged.

ACKNOWLEDGMENTS

DF-C thanks Xunta de Galicia for financial support through a postdoctoral grant.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2020.00016/full#supplementary-material

It contains the MS-QH and MS-T(C) partition functions, as well as the electronic energy and the Cartesian coordinates of all conformations for the species ranging from n-propanol to n-heptanol. For L-serine, Cartesian coordinates of the 60 HL conformers are also listed, as well as a figure labeling the targeted torsions.

REFERENCES

Alecu, I. M., Zheng, J., Zhao, Y., and Truhlar, D. G. (2010). Computational thermochemistry: scale factor databases and scale factors for vibrational frequencies obtained from electronic model chemistries. *J. Chem. Theory Comput.* 6, 2872–2887. doi: 10.1021/ct100326h

Black, G., and Simmie, J. M. (2010). Barrier heights for H-stom abstraction by HO_2 from n-butanol –A simple yet exacting test for model chemistries? *J. Comput. Chem.* 31, 1236–1248. doi: 10.1002/jcc.21410

Chen, L., Zhu, W., Lin, K., Hu, N., Yu, Y., Zhou, X., et al.(2015). Identification of alcohol conformers by Raman spectra in the C-H stretching region. *J. Phys. Chem. A* 119, 3209–3217. doi: 10.1021/jp513027r

Fernández-Ramos, A. (2013). Accurate treatment of two-dimensional non-separable hindered internal rotors. *J. Chem. Phys.* 138:134112. doi: 10.1063/1.4798407

Fernández-Ramos, A., Ellingson, B. A., Meana-Pañeda, R., Marques, J. M. C., and Truhlar, D. G. (2007). Symmetry numbers and chemical reaction rates. *Theor. Chem. Acc.* 118, 813–826. doi: 10.1007/s00214-007-0328-0

Ferro-Costas, D., Cordeiro, M. N. D. S., Truhlar, D. G., and Fernández-Ramos, A. (2018a). Q2dtor: a program to treat torsional anharmonicity through coupled pair of torsions in flexible molecules. *Comput. Phys. Commun.* 232, 190–205. doi: 10.1016/j.cpc.2018.05.025

Ferro-Costas, D., Martínez-Núńez, E., Rodríguez-Otero, J., Cabaleiro-Lago, E., Estévez, C. M., Fernández, B., et al.(2018b). Influence of multiple conformations and paths on rate constants and product branching ratios. thermal decomposition of 1-propanol radicals. *J. Phys. Chem. A* 122, 4790–4800. doi: 10.1021/acs.jpca.8b02949

Friedrich, N.-O., Flachsenberg, F., Meyder, A., Sommer, K., Kirchmair, J., and Rarey, M. (2019). Conformator: a novel method for the generation of conformer ensembles. *J. Chem. Inf. Model.* 59, 731–742. doi: 10.1021/acs.jcim.8b00704

Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., et al.(2004). *Gaussian 09, Revision A.02*. Wallingford, CT: Gaussian, Inc.

Hehre, W. J., Ditchfield, R., and Pople, J. A. (1972). Self-consistent molecular orbital methods. XII. Further extensions of Gaussian-type basis sets for use in molecular orbital studies of organic molecules. *J. Chem. Phys.* 56, 2257–2261. doi: 10.1063/1.1677527

Klippenstein, S. J. (2017). From theoretical reaction dynamics to chemical modeling of combustion. *Proc. Combust. Inst.* 36, 77–111. doi: 10.1016/j.proci.2016.07.100

Kolossváry, I., and Guida, W. C. (1996). Low mode search. an efficient, automated computational method for conformational analysis: application to cyclic and acyclic alkanes and cyclic peptides. *J. Am. Chem. Soc.* 118, 5011–5019. doi: 10.1021/ja952478m

Loferer, M. J., Kolossváry, I., and Aszódi, A. (2007). Analyzing the performance of conformational search programs and compound databases. *J. Mol. Graph. Model.* 25, 700–710. doi: 10.1016/j.jmgm.2006.05.008

Najbauer, E. E., Bazsó G., Apóstolo, R., Fausto, R., Biczysko, M., Barone, V., et al.(2015). Identification of serine conformers by matrix-Isolation IR sspectroscopy aided by near-infrared laser-induced conformational change, 2D correlation analysis, and quantum mechanical anharmonic computations. *J. Phys. Chem. B* 119, 10496–10510. doi: 10.1021/acs.jpcb.5b05768

O'Boyle, N. M., Vandermeersch, T., Flynn, C. J., Maguire, A. R., and Hutchison, G. R. (2011). Confab - systematic generation of diverse low-energy conformers. *J. Cheminformat.* 3:8. doi: 10.1186/1758-2946-3-8

Simón-Carballido, L., Bao, J. L., Alves, T. V., Meana-Pañeda, R., Truhlar T. G., and Fernández Ramos, A. (2017). Anharmonicity of coupled torsions: the extended two-dimensional torsion method and its use to assess more approximate methods. *J. Chem. Theory Comput.* 13, 3478–3492. doi: 10.1021/acs.jctc.7b00451

Vaskivskyi, Y., Chernolevska, Y., Vasylieva, A., Pogorelov, V., Pratakyte, R., Stocka, J., et al.(2019). 1-Hexanol conformers in a nitrogen matrix: FTIR study and high-level ab initio calculations *J. Mol. Liq.* 278, 356–362. doi: 10.1016/j.molliq.2019.01.059

Watts, K. S., Dalal, P., Murphy, R. B., Sherman, W., Friesner, R. A., and Shelley, J. C. (2010). Confgen: a conformational search method for efficient generation of bioactive conformers. *J. Chem. Inf. Model.* 40, 534–546. doi: 10.1021/ci100015j

Yu, T., Zheng, J., and Truhlar, D. G. (2011). Multi-structural variational transition state theory. Kinetics of the 1,4-hydrogen shift isomerization of the pentyl radical with torsional anharmonicity. *Chem. Sci.* 2, 2199–2213. doi: 10.1039/c1sc00225b

Zhao, Y., Lynch, B. J., and Truhlar, D. G. (2004). Hybrid meta density functional theory methods for thermochemistry, thermochemical kinetics, and noncovalent interactions: the MPW1B95 and MPWB1K models and comparative assessments for hydrogen bonding and van der Waals interactions. *J. Phys. Chem A* 108, 6908–6918. doi: 10.1021/jp048147q

Zheng, J., Meana-Pañeda, R., and Truhlar, D. G. (2013). Mstor version 2013: a new version of the computer code for the multi-structural torsional anharmonicity, now with a coupled torsional potential. *Comput. Phys. Commun.* 184, 2032–2033. doi: 10.1016/j.cpc.2013.03.011

Zheng, J., Mielke, S. L., Clarkson, K. L., and Truhlar, D. G. (2012). Mstor: a program for calculating partition functions, free energies, enthalpies, entropies, and heat capacities of complex molecules including torsional anharmonicity. *Comput. Phys. Commun.* 183, 1803–1812. doi: 10.1016/j.cpc.2012.03.007

Zheng, J., and Truhlar, D. G. (2013). Quantum thermochemistry: multistructural method with torsional anharmonicity based on a coupled torsional potential. *J. Chem. Theory Comput.* 9, 1356–1367. doi: 10.1021/ct3010722

Zheng, J., Yu, T., Papajak, E., Alecu, I. M., Mielke, S. L., and Truhlar, D. G. (2011a). Practical methods for including torsional anharmonicity in thermochemical calculations on complex molecules: the internal-coordinate multi-structural approximation. *Phys. Chem. Chem. Phys.* 13, 10885–10907. doi: 10.1039/c0cp02644a

Zheng, J., Yu, T., and Truhlar, D. G. (2011b). Multi-structural thermodynamics of C-H bond dissociation in hexane and isohexane yielding seven isomeric hexyl radicals. *Phys. Chem. Chem. Phys.* 13, 19318–19324. doi: 10.1039/c1cp21829h

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS

Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION

Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW

Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW

Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH

Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING

Articles designed
for optimal readership
across devices

FOLLOW US

@frontiersin

IMPACT METRICS

Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION

Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK

Our network
increases your
article's readership