frontiers in Genetics

TECHNOLOGY AND CODE
published: 30 October 2019
doi: 10.3389/fgene.2019.00990

ContraDRG: Automatic Partial Charge Prediction by Machine Learning

*Roman Martin [1,2] and Dominik Heider [1]**

[1] Department of Mathematics and Computer Science, University of Marbug, Marburg, Germany, [2] Department of Organic-Analytical Chemistry, TUM Campus Straubing, Straubing, Germany

In recent years, machine learning techniques have been widely used in biomedical research to predict unseen data based on models trained on experimentally derived data. In the current study, we used machine learning algorithms to emulate computationally complex predictions in a reverse engineering–like manner and developed ContraDRG, a software that can be used to predict partial charges for small molecules based on PRODRG and Automated Topology Builder (ATB) predictions. Both tools generate molecular topology files, including the partial atomic charge, by using different procedures. We show that ContraDRG can accurately predict partial charges in a fraction of the time, because it exploits existing complex models with intensive calculations by using machine learning techniques and thus can also be applied for screening projects with large amounts of molecules. We provide ContraDRG as a web server, which can be used to automatically assign partial charges to incoming user-specified molecules by using our machine learning models. In this study, we compared ContraDRG with PRODRG and ATB in regard of predictivity by statistical methods. ContraDRG allows predicting ATB-derived partial charges with an R^2 value up to 0.980 and for PRODRG up to 1.00. While ATB requires hours or days for the quantum mechanical accurate calculation and refinements, ContraDRG does its approximation within seconds.

Keywords: PRODRG, ATB, machine learning, molecular dynamics simulations, partial charge prediction

OPEN ACCESS

Edited by:
Harinder Singh,
J. Craig Venter Institute,
United States

Reviewed by:
Kumardeep Chaudhary,
Icahn School of Medicine at
Mount Sinai,United States
Sandeep Kumar Dhanda,
La Jolla Institute for Immunology (LJI),
United States

****Correspondence:***
Dominik Heider
dominik.heider@uni-marburg.de

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *26 June 2019*
Accepted: *18 September 2019*
Published: *30 October 2019*

Citation:
Martin R and Heider D (2019) ContraDRG: Automatic Partial Charge Prediction by Machine Learning. Front. Genet. 10:990. doi: 10.3389/fgene.2019.00990

INTRODUCTION

In the last decades, several studies demonstrated how machine learning algorithms were able to create accurate predictions or classifications from experimentally derived data. The applications of machine learning algorithms in biomedical research are diverse (Larrañaga et al., 2006) and range from single-molecule interaction prediction for drug design (Lavecchia, 2015) or omics pattern recognition (Stanke and Morgenstern, 2005), toward the prediction of entire biological systems (D'Alche-Buc and Wehenkel, 2008).

However, in the current study, we used machine learning algorithms to emulate computationally intensive calculations. Precise determination of topology parameters for small molecules, particularly partial charges, is a crucial step for molecular dynamics (MD) simulations and other biochemical and biophysical computations. In particular, MD simulations depend heavily on the accurate parameterization of the molecules; otherwise, the simulations tend to be unreliable and misleading (Lemkul et al., 2010). One main challenge for generating reliable predictions is the ability to create a force field consistent topology for new small molecules since the force fields theory is mostly derived from empirical analysis.

For this purpose, there are different force fields available, based on diverse parameters and underlying theories, such as GROMOS (van Gunsteren et al., 1996; Daura et al., 1998; Scott et al., 1999; Schuler et al., 2001; Oostenbrink et al., 2004), OPLS (Jorgensen and Tirado-Rives, 1988; Jorgensen et al., 1996), CHARMM (Patel and Brooks, 2004; Patel et al., 2004), and AMBER (Cornell et al., 1995; Wang et al., 2004). Parameterization for synthetic small molecules is supported by the general AMBER force field (Wang et al., 2004) and the general CHARMM force field (Patel and Brooks, 2004; Patel et al., 2004), in contrast to GROMOS and OPLS. While detailed information about the GROMOS96 parameter sets is not publicly available, OPLS-AA reveals their entire parameter sets, which includes geometry optimization and quantum chemical calculations (Jorgensen et al., 1996; Kaminski et al., 2001). Thus, users of the GROMOS96 force field rely on empirical parameters and subsequent validations by thermodynamic integration (Oostenbrink et al., 2004).

Over the last years, some freely available tools were developed, refined, and established for automated topology generation. Two commonly used tools are PRODRG (Van Aalten, 1996; Schüttelkopf and Van Aalten, 2004) and the Automated Topology Builder (ATB) (Malde et al., 2011; Koziara et al., 2014; Stroet et al., 2018). Both are frequently used tools that receive user-defined small-molecule files and return parameterized GROMOS-compatible topology files including their partial atomic charges. While PRODRG partial charge determination is based on mapping of building blocks and charge groups onto a database, ATB uses quantum chemical calculations involving electron densities and geometry optimizations (Chandra Singh and Kollman, 1984). However, PRODRG is much faster compared to ATB and produces topologies within seconds, while ATB requires up to multiple days, but generates more precise, more reliable, and more consistent results (Lemkul et al., 2010; Malde et al., 2011). Both tools have been already used for protein–peptide, protein–ligand, protein–lipid, and pharmaceutical drug optimizations (Santos et al., 2017). Although both tools provide free access for automated file parameterization, only ATB supplies a modern application programming interface. Additionally, there are several stand-alone tools, such as Open Babel (O'Boyle et al., 2011) and AutoDock Tools (Morris et al., 2009), which can predict partial charges based on different methods like MMFF94 (Halgren, 1999), based on quantum chemical calculations, or QTPIE (Chen and Martı, 2007), which describes the flow in molecules based on charge transfer variables.

While PRODRG and ATB are proprietary software, they do provide free access for academic purpose. Contrary to that, fully proprietary software like VeraChem's VCharge or Schroedinger's Maestro, which predict, among others, partial charges are also available. VCharge uses a method based on QM-derived electronegativity equalization (Gilson et al., 2003), and Maestro computes the charges according CM1A-BCC (OPLS3e) (Marenich et al., 2012; Roos et al., 2019). Additionally, there is proprietary software such as Amber that requires external tools for partial charges predictions, like the provided and recommended free antechamber (Wang et al., 2006). Antechamber applies usually the AM1-BCC method (Jakalian et al., 2002) for small molecules and can be optimized with provided QM calculations by the RESP method (Bayly et al., 1993).

Engler et al. (2019) showed recently in an innovative approach how to solve two common problems of partial charge determination: (i) the single partial-charge assignment per atom and (ii) the total charge determination. By transferring these problems into a multiple-choice knapsack problem (Dudziński and Walukiewicz, 1987; Kellerer et al., 2004), they were able to predict the partial charges automatically. Moreover, a recent study showed that machine learning prediction based on quantum-chemical calculation can be used to predict partial charges (Bleiziffer et al., 2018).

In the current study, we used small-molecule three-dimensional structures files for prediction of partial charges, based on machine-derived data from the web tools PRODRG and ATB. To this end, we analyzed and compared a set of different machine learning methods and emulated the aforementioned tools. Finally, we compared our predictions with the existing tools. This study demonstrates the usefulness of machine learning models for reverse engineering of costly calculations, which are provided in an easy-to-use online tool.

MATERIALS AND METHODS

Dataset

This study is based on two different datasets, namely, the PRODRG dataset and the ATB dataset. The PRODRG dataset is based on randomly selected molecule structures from the PubChem database (Kim et al., 2018). These molecules were converted into Protein Database Bank format *via* Open Babel (O'Boyle et al., 2011) and subsequently predicted *via* the PRODRG server (v. AA100323.0717). Energy minimization was deactivated, and full charge prediction and chirality enabled. The ATB dataset was collected from the curated molecule and topology files from the ATB (v. 3.0) database (Stroet et al., 2018). We mapped the partial charge predictions from the topology files with the provided all-atom Protein Database Bank files.

We calculated the pairwise Tanimoto similarity coefficient *via* Open Babel (linear seven atoms fragments) for all files to ensure that a diverse set of molecules was used (Kim et al., 2018). The Tanimoto coefficient represents a known indicator for molecular structure similarities (Bajusz et al., 2015). Therefore, we determined the coefficient by comparing every molecule to each other. The resulting coefficients were drawn into a violin plot.

Feature Encoding

In the current study, we focused only on organic elements, namely, carbon, hydrogen, nitrogen, oxygen, phosphorus, sulfur, fluorine, bromine, and iodine (C, H, N, O, P, S, F, Cl, Br, and I). We used 61 different features for the encoding of the molecules, where all atoms are individually analyzed (**Figure 1**). Molecules are internally represented as a cyclic undirected graph, where atoms correspond to vertices, and bonds to edges. These

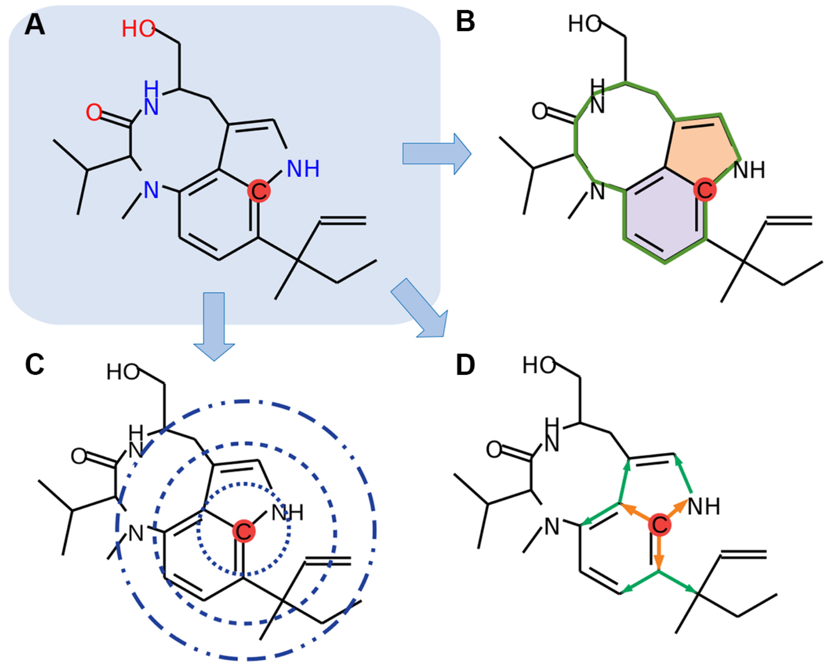

FIGURE 1 | Schematic overview of the feature encoding. **(A)** Each atom will be selected (red dot), and encodings will be generated **(B–D)**. **(B)** Overall circular structures (green line) and nested (colored areas) are detected by a depth-first search. **(C)** Distance searches with three different radii are applied. **(D)** Second-level neighbors path tracing is implemented (orange arrows, first level; green arrows, second level). Chemical structures were drawn with MolView (https://molview.org).

encodings include the hybridization state of carbon atoms, sizes and amounts of nested circles, distances to adjacent atoms, and presence of neighbors through a second-level path tracing. Nested circular structures were identified by a depth-first search derived from the graph theory.

To encode an entire molecule, a list of the positions of the atoms and an adjacency matrix for the bonds are necessary. Protein Database Bank files and SMILE (Weininger, 1988) files can be encoded in such a way easily. However, in contrast to existing approaches, we take explicitly the three-dimensional information into account, thus allowing making prediction also for theoretical molecules.

Machine Learning

We used the R package caret (v. 6.0-81) (Max and Kuhn, 2008) for building the machine learning models. We build models for each element independently. The datasets (one dataset for each element) were split into train and test data with a ratio of 1:4. We trained different models including linear regression, stochastic gradient boosting (Friedman, 2002), random forests (RF) (Breiman, 2001), quantile regression forests (Meinshausen, 2006), weighted k-nearest neighbors (Altman, 1992), and support vector machines (SVMs) (Cortes and Vapnik, 1995) with different kernels. RFs were trained with 500 trees and k-nearest neighbors were built based on $k = 7$ and a Minkowski distance of 2. All other models were trained with default parameters. All models were trained with the partial charge values as labels from PRODRG or ATB, respectively. The models are evaluated based on root median square error (RMSE):

$$RMSE = \sqrt{\frac{\Sigma_{T}^{t=1}(\hat{y}_t - y_t)^2}{T}} \tag{1}$$

Furthermore we used the normalized RMSE:

$$NRMSE = \frac{\sqrt{\frac{\Sigma_{T}^{t=1}(\hat{y}_t - y_t)^2}{T}}}{\sqrt{(min(y) - max(y))^2}} \cdot 100 \tag{2}$$

A direct comparison between the different software tools, respectively, the algorithms, is not possible since the applications are using different force fields. However, the aforementioned metrics enable a direct comparison of the machine learning predictions to the original software.

Molecular Dynamics

We tested the ATB-derived random forest models, with 50 randomly chosen molecules from the ATB database with experimental hydration free energy (ΔG^{hyd}). Topologies and coordinate files were obtained by the ATB database. Parameters for the molecule dynamics were taken from the FreeSolv database (Mobley et al., 2009; Mobley, 2013; Mobley and Guthrie, 2014; Duarte Ramos Matos et al., 2017). We used the *gromos54a7_atb.ff* force field according to ATB. Simulations were run under GROMACS (v. 2016.3) with NPT conditions at 298 K and 1 atm. The cutoff for the van der Waals (rvdw) and electrostatic interactions (rcoulomb) was set to

1.2 nm. The simulations were performed with 20 λ-steps and 2 fs per time step, resulting in 12.5-ns simulations per λ-point. GROMACS simulations require removing all nonpolar hydrogens for a united-atoms model. For ContraDRG, original partial charges from ATB were overwritten with ContraDRG predictions. Therefore, we summarized all removed charges into the adjacent remaining atom. Atom-centered partial charge predictions occasionally generate molecules with an excess of net total charges. The excess was eliminated by distributing the excess equally through a molecule. A comparison of the absolute errors between the experimental ΔG^{hyd} free energy and ATB and that between the experimental ΔG^{hyd} free energy with ContraDRG were performed by a Welch *t* test (Welch, 1947). We omitted MD simulations with PRODRG topologies since it has been reported as inaccurate (Lemkul et al., 2010), which could be confirmed in our analyses.

Web Application

The web application ContraDRG is based on an Apache web server (v. 2.4.29) with PHP (v. 7.2.17) and R (v. 3.4.4) as background services. Incoming data will be filtered and converted by Open Babel (v. 2.4.1) into temporary internal PDB files. ContraDRG reads the PDB structures, performs the feature encoding, and applies the trained machine learning models. The final output will be generated by the Open Babel and remapped with partial charge values predicted by ContraDRG determining partial charge values. A two-dimensional graph of the molecule will be displayed after the machine learning prediction. Missing three-dimensional molecules structures, as provided by SMILES formatted molecules, will be computed by Open Babel as well. The partial charge prediction will be performed by the random forest models for each element, which have been shown to outperform the other models.

RESULTS

Overall Approach

The current study aimed to build a reliable and fast prediction model for partial charges. To this end, we used machine learning algorithms to emulate computationally complex predictions in a reverse engineering–like manner and developed ContraDRG, a software that can be used to predict partial charge assignments based on PRODRG and ATB predictions. We collected thousands of randomly selected molecules from PubChem and the ATB database. Finally, we provide the freely accessible web tool ContraDRG, which can be used for partial charge predictions. The resulting predictions provide a reliable approximation of the original tools. However, predictions are carried out in seconds without any user restrictions.

Datasets

We collected 7,000 molecule structures from PubChem with an average size of 19 heavy atoms per molecule (resulting in 132,859 atoms), which were predicted using PRODRG. Seventy percent of the atoms in the PRODRG dataset are carbon, and 13% are oxygen atoms. Moreover, we randomly collected 10,000 molecules from the ATB database with an average size of 25 heavy atoms per molecule. In this ATB dataset, 47% of the atoms are hydrogens, while 35% are carbons. **Figure 2** represents the distribution of all elements in our datasets. Variances in the number of hydrogen atoms between both datasets are due to differences in the underlying model, namely, united-atoms model for PRODRG and all-atoms model for ATB.

To achieve a high variety of different molecules, we analyzed the similarities between every molecule structure to each other by calculating the Tanimoto coefficient in a pairwise manner. The

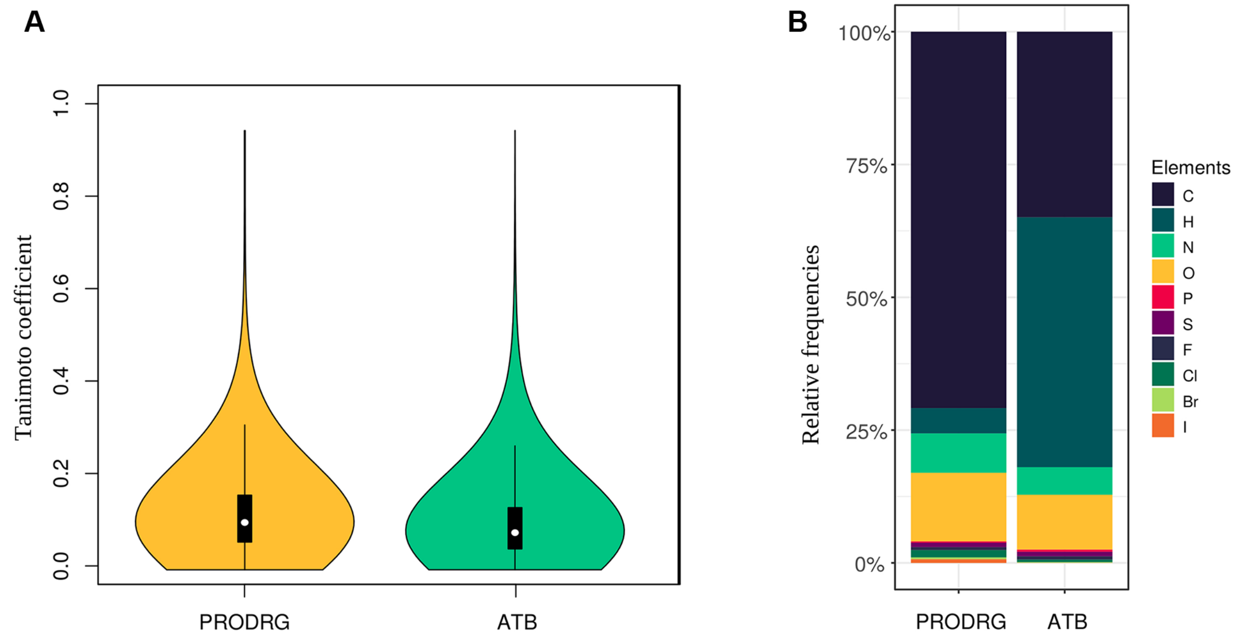

FIGURE 2 | (A) The violin plots show the Tanimoto coefficient for both datasets. The plot width correlates with the relative frequencies of the coefficient. The white dot represents the median, while the black box represents the interquartile range, and the black lines, the 95% confidence intervals. One-sample *t* tests for both sets of Tanimoto coefficients show a $p < 0.001$ for a mean below 0.15. **(B)** The distribution of atom types for each dataset is represented by relative bar plots.

Tanimoto coefficients and their distribution for the PRODRG and the ATB datasets are shown as violin plots in **Figure 2**. The coefficients of all possible pairs of molecules are relatively low, with a median of around 0.11 for the PRODRG and 0.08 for the ATB dataset, indicating a high variance between the incorporated molecules. We used a one-sample *t* test on the Tanimoto coefficients for testing significance against a mean value of 0.15 ($p < 0.001$).

Analysis of the charge distribution through all elements shows a variance in the charge predictions between the different datasets in **Figure 3**. Since the occurrence of molecular constitutions and conformations is limited, the partial charges are not equally distributed over the whole range. Moreover, some atoms tend to act as an electron-pair donor, such as oxygen. Therefore, most oxygen is charged negatively or neutral. Generally, the charge predictions differentiate heavily between the PRODRG and ATB datasets. PRODRGs predictions are more clustered than ATB. This clustering can be observed in the shape of the charge distribution curves by the present peaks of the PRODRG dataset in **Figure 3**. One explanation for the highly clustered charges of PRODRG is the fact that PRODRG maps the molecule to a limited set of building blocks and charge groups, while ATB refines partial charges after an initial determination according to the Merz–Singh–Kollman method (Chandra Singh and Kollman, 1984).

Partial Charge Prediction

We employed several machine learning algorithms for every element on each dataset. Depending on the number of data points, the machine learning algorithm training took several hours up to 10 days on a high-performance cluster, especially for the SVMs and random forest models. Linear regression models turned out to be most inaccurate compared to the random forest models, which mostly outperform all other models in both datasets. For this reason, the ContraDRG web application uses random forest models for the prediction. An exemplary direct side-by-side comparison of ATB-derived ContraDRG prediction with ATB 3.0 is provided in the **Supplementary Material**. For a set of 50 randomly chosen molecules, ATB required an average execution time of 8 h for generating the topology including the partial charges, while ContraDRG required only 9.2 seconds on average for the partial charge prediction per molecule.

Table 1 represents a shortened overview of the best prediction performance. The full-length table is provided in the **Supplementary Material**. The normalized RMSE values allow an easy comparison for each element since they are normalized to the whole range of present partial charge values. Moreover, the predictions for PRODRG-derived data are more accurate than for ATB, which can be observed particularly for underrepresented elements such as iodine in the ATB dataset. The mean R^2 for PRODRG predictions is 0.962 (min. 0.791, max. 1.000) for random forest and 0.685 (0.010–0.985) for SVMs with linear kernel in comparison to the ATB predictions with a mean of R^2 0.908 (0.778–0.982) for random forest and 0.744 (0.520-0.971) for linear SVMs. Overall, the predictions based on the random forest models are more accurate than those based on the other models.

The MD analyses show that the predictions of ContraDRG's ATB-derived random forest models perform as well as ATB in terms of the ΔG^{hyd} free energy calculation. Furthermore, we compared the errors between experimental ΔG^{hyd} values and those derived from ATB with the errors between the experimental data and ATB-derived ContraDRG prediction. No significant differences have been observed by using the Welch *t* test ($p = 0.53$) (Max and Kuhn, 2008). Additional information is provided as **Supplementary Materials**.

DISCUSSION

In summary, we were able to produce partial charge predictions by our fast and unrestricted approach. Depending on the dataset and the frequency of an element in the dataset, reliable predictions are possible. The models for underrepresented elements such as chlorine, bromine, and iodine performed worse compared to those trained on the most abundant elements such as carbon or hydrogen. Surprisingly, linear regression performed better for iodine in the ATB dataset than the corresponding random forest model (see **Supplementary Material**). A possible explanation for that is the fact that iodine atoms are the most underrepresented elements in the ATB dataset, and the random forest models tend to overfit.

Generally, as **Table 1** shows, our predictions for the PRODRG dataset are more accurate than for ATB. There are several possible reasons for that. First, PRODRG is based on a simpler method for assigning partial charges (Altman, 1992). Second, we used molecules from the PubChem database for the PRODRG dataset. The three-dimensional structures of these molecules are all idealized and normalized by PubChem (Bolton et al., 2008). Compared to that, we used curated molecules for the ATB dataset, which mostly originate from the manually curated ChEMBL database (Gaulton et al., 2012; Stroet et al., 2018). Third, ATB performs geometric optimization and remaps the partial charges back to the original structures. Geometry-optimized charges cannot be learned by our model since we do not take geometrical temporary changes into account. Additionally, as shown in **Figure 3**, the partial charges for the ATB data have a higher variance, which makes prediction generally more difficult.

Although our approach is biased to inherit errors from the original tools, the predictions achieve a reliable approximation with low RMSE values. Inconsistent partial charges, which can appear in PRODRG (Lemkul et al., 2010), are unlikely because our models predict the charges along with defined models without determinations of building blocks. Error propagation cannot be avoided; however, by using larger datasets and extended feature sets, the prediction models tend to be more accurate. Our web tool is freely accessible at http://contradrg.heiderlab.de.

CONCLUSION

All existing approaches of partial charges predictions for molecules aim at reconstructing the exact empirical-validated value. Thus, the computations are based on empirical determined data (Mortier et al., 1986; Besler et al., 1990) or on quantum

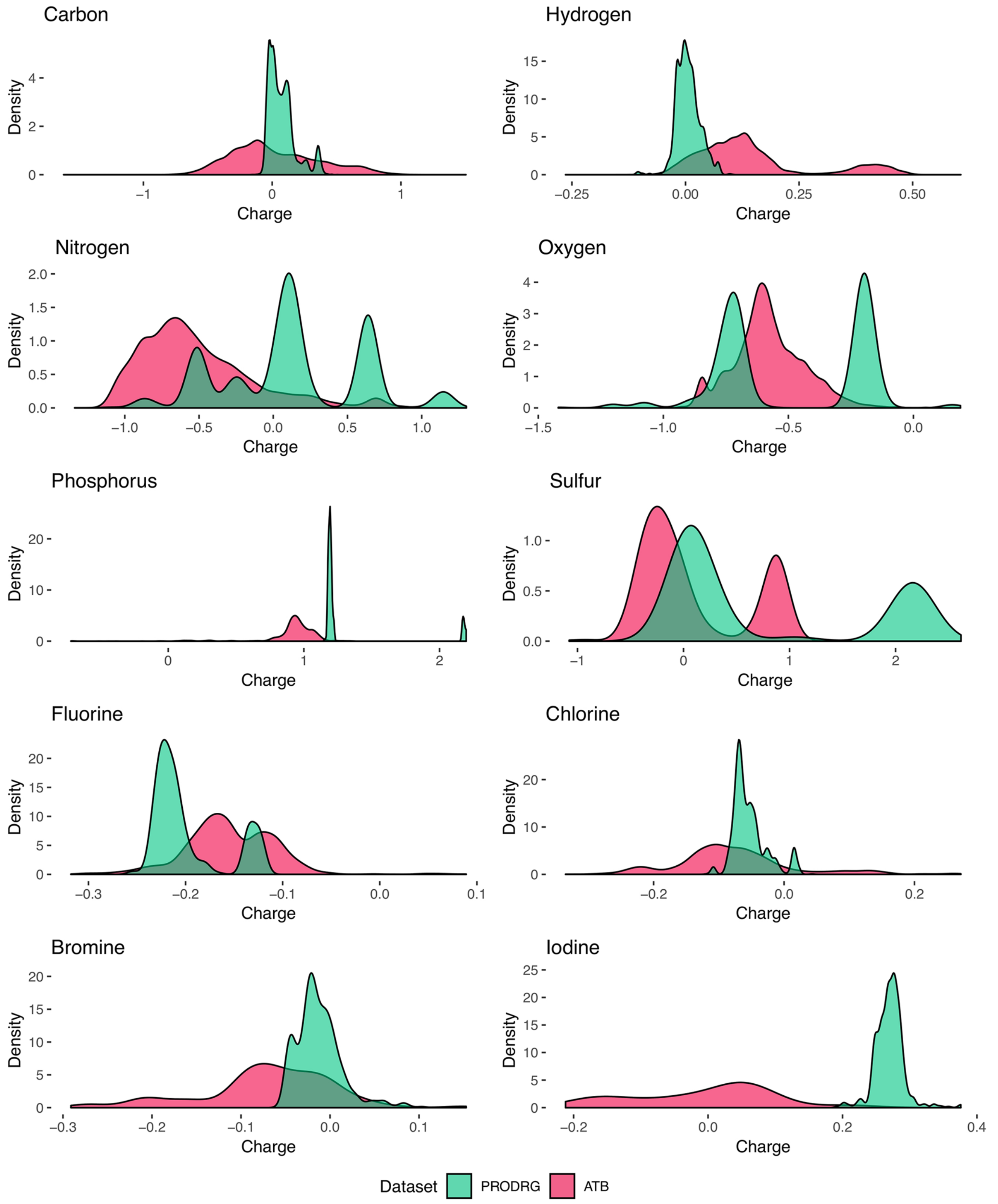

FIGURE 3 | Smoothed kernel density estimates represent the distribution of partial charges (units of *e*) for each molecule in the datasets. Distribution from PRODRGs dataset reveals more clustered peaks (green) than from ATB (red).

TABLE 1 | Performance comparison for partial charge prediction (units of *e*) by random forest and support vector machines with linear kernel of the PRODRG and ATB dataset.

	PRODRG						ATB					
	Random forest			SVM linear			Random forest			SVM linear		
	RMSE	NRMSE	R^2	RMSE	NRMSE	R^2	RMSE	NRMSE	R^2	RMSE	NRMSE	R^2
C	0.011	1.443	0.989	0.054	7.073	0.738	0.069	2.398	0.961	0.152	5.268	0.810
H	0.005	2.878	0.955	0.026	13.924	0.010	0.018	2.313	0.980	0.046	5.794	0.879
N	0.048	1.986	0.990	0.249	10.374	0.730	0.113	5.391	0.919	0.163	7.772	0.834
O	0.051	3.184	0.971	0.153	9.494	0.739	0.047	4.200	0.887	0.071	6.302	0.746
P	0.002	0.152	1.000	0.073	7.157	0.965	0.075	3.712	0.892	0.097	4.803	0.823
S	0.015	0.678	1.000	0.120	5.454	0.985	0.068	3.095	0.982	0.087	3.962	0.971
F	0.003	2.436	0.993	0.007	5.184	0.968	0.017	4.179	0.897	0.037	9.205	0.520
Cl	0.004	2.724	0.980	0.020	15.293	0.415	0.030	5.490	0.895	0.054	9.796	0.705
Br	0.011	8.625	0.791	0.016	12.222	0.589	0.033	8.796	0.778	0.049	13.033	0.531
I	0.004	2.575	0.955	0.010	6.592	0.706	0.036	12.840	0.888	0.062	22.082	0.624
$\bar{x}$)	0.015	2.668	0.962	0.073	9.277	0.685	0.051	5.241	0.908	0.082	8.802	0.744

The root median square error (RMSE) represents the quality of errors while NRMSE shows a normalized RMSE.

mechanical theories (Manz and Sholl, 2010; Manz and Sholl, 2012; Manz and Limas, 2016). However, our approach tries to emulate the algorithm of the predictor without implementing any background knowledge about the underlying theories. Analysis of the input and output data from the web servers with subsequent machine learning approaches are sufficient to easily compute reliable approximations. Our web tool can be used to assign partial charge predictions automatically within seconds. This allows, for example, the correction of precalculated topology files. In the future, we intend to improve our models by using more training data, in particular for those atoms that are underrepresented, and to extend the feature set. Additionally, we intend to generate GROMOS-compatible topology files without geometrical optimization for molecular dynamics simulations.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in http://cdrg.mathematik.uni-marburg.de/data/raw-dataset.zip.

AUTHOR CONTRIBUTIONS

RM performed the data and machine learning analysis. RM drafted the manuscript. DH supervised the project, discussed the results, and revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This study was funded by the European Regional Development Fund, EFRE-Program, European Territorial Cooperation (ETZ) 2014-2020, Interreg V A, Project 41.

ACKNOWLEDGMENTS

We thank the ATB team, particularly Martin Stroet and Alan Mark for providing us access to the ATB database. Calculations were carried out on the MaRC2 high performance cluster of the University of Marburg, which are supported by the State Ministry of Higher Education, Research and the Arts. We especially thank Mr. Sitt of HPC-Hessen for his technical support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00990/full#supplementary-material

REFERENCES

Altman, N. S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* 46, 175. doi: 10.2307/2685209

Bajusz, D., Rácz, A., and Héberger, K. (2015). Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminf.* 7, 1–13. doi: 10.1186/s13321-015-0069-3

Bayly, C. I., Cieplak, P., Cornell, W., and Kollman, P. A. (1993). A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* 97, 10269–10280. doi: 10.1021/j100142a004

Besler, B. H., Merz, K. M., and Kollman, P. A. (1990). Atomic charges derived from semiempirical methods. *J. Comput. Chem.* 11, 431–439. doi: 10.1002/jcc.540110404

Bleiziffer, P., Schaller, K., and Riniker, S. (2018). Machine Learning of Partial Charges Derived from High-Quality Quantum-Mechanical Calculations. *J. Chem. Inf. Model.* 58, 579–590. doi: 10.1021/acs.jcim.7b00663

Bolton, E. E., Wang, Y., Thiessen, P. A., and Bryant, S. H. (2008). *Chapter 12 PubChem: Integrated Platform of Small Molecules and Biological Activities* Vol. 4. Elsevier B.V, 217–241. Amsterdam, Netherlands. doi: 10.1016/S1574-1400(08)00012-1

Breiman, L. (2001). Random Forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Chandra Singh, U., and Kollman, Peter A (1984). An approach to computing electrostatic charges for molecules. *J. Comput. Chem.* 5, 129–145. doi: 10.1002/jcc.540050204

Chen, J., and Martı, T. J. (2007). QTPIE: Charge transfer with polarization current equalization. A fluctuating charge model with correct asymptotics. *Chem. Physics Letters* 438, 315–320. doi: 10.1016/j.cplett.2007.02.065

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., et al. (1995). A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* 117, 5179–5197. doi: 10.1021/ja00124a002

Cortes, C., and Vapnik, V. (1995). Support-Vector Networks. *Mach. Learn.* 20, 273–297. doi: 10.1.1.170.5707

D'Alche-Buc, F., and Wehenkel, L. (2008). Machine learning in systems biology. *BMC Proc.* 2 Suppl 4, S1. doi: 10.1186/1753-6561-2-s4-s1

Daura, X., Mark, A. E., and Van Gunsteren, W. F. (1998). Parametrization of aliphatic CHn united atoms of GROMOS96 force field. *J. Comput. Chem.* 19, 535–547. doi: 10.1002/(SICI)1096-987X(19980415)19:5⟨535::AID-JCC6⟩3.0.CO;2-N

Duarte Ramos Matos, G., Kyu, D. Y., Loeffler, H. H., Chodera, J. D., Shirts, M. R., and Mobley, D. L. (2017). Approaches for Calculating Solvation Free Energies and Enthalpies Demonstrated with an Update of the FreeSolv Database. *J. Chem. Eng. Data* 62, 1559–1569. doi: 10.1021/acs.jced.7b00104

Dudziński, K., and Walukiewicz, S. (1987). Exact methods for the knapsack problem and its generalizations. *Eur. J. Oper. Res.* 28, 3–21. doi: 10.1016/0377-2217(87)90165-2

Engler, M. S., Caron, B., Veen, L., Geerke, D. P., Mark, A. E., and Klau, G. W. (2019). Automated partial atomic charge assignment for drug-like molecules: a fast knapsack approach. *Algorithms Mol. Biol.* 14, 1. doi: 10.1186/s13015-019-0138-7

Friedman, J. H. (2002). Stochastic gradient boosting. *Comput. Stat. Data Anal.* 38, 367–378. doi: 10.1016/S0167-9473(01)00065-2

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., et al. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40, D1100–D1107. doi: 10.1093/nar/gkr777

Gilson, M. K., Gilson, H. S., and Potter, M. J. (2003). Fast Assignment of Accurate Partial Atomic Charges: An Electronegativity Equalization Method that Accounts for Alternate Resonance Forms. *J. Chem. Inf. Comput. Sci.* 43, 1982–1997. doi: 10.1021/ci034148o

Halgren, T. A. (1999). MMFF VI. MMFF94s option for energy minimization studies. *J. Comput. Chem.* 20, 720–729. doi: 10.1002/(SICI)1096-987X(199905)20:7⟨720::AID-JCC7⟩3.0.CO;2-X

Jakalian, A., Jack, D. B., and Bayly, C. I. (2002). Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.* 23, 1623–1641. doi: 10.1002/jcc.10128

Jorgensen, W. L., and Tirado-Rives, J. (1988). The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* 110, 1657–1666. doi: 10.1021/ja00214a001

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

Kaminski, G. A., Friesner, R. A., Tirado-Rives, J., and Jorgensen, W. L. (2001). Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins *via* Comparison with Accurate Quantum Chemical Calculations on Peptides †. *J. Phys. Chem. B* 105, 6474–6487. doi: 10.1021/jp003919d

Kellerer, H., Pferschy, U., and Pisinger, D. (2004). *Knapsack Problems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 548. doi: 10.1007/978-3-540-24777-7

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., et al. (2018). PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* 47, 1–8. doi: 10.1093/nar/gky1033

Koziara, K. B., Stroet, M., Malde, A. K., and Mark, A. E. (2014). Testing and validation of the Automated Topology Builder (ATB) version 2.0: Prediction of hydration free enthalpies. *J. Comput.-Aided Mol. Design* 28, 221–233. doi: 10.1007/s10822-014-9713-7

Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., et al. (2006). Machine learning in bioinformatics. *Briefings Bioinf.* 7, 86–112. doi: 10.1093/bib/bbk007

Lavecchia, A. (2015). Machine-learning approaches in drug discovery: methods and applications. *Drug Discovery Today* 20, 318–331. doi: 10.1016/j.drudis.2014.10.012

Lemkul, J. A., Allen, W. J., and Bevan, D. R. (2010). Practical Considerations for Building GROMOS-Compatible Small Molecule Topologies. *J. Chem. Inf. Model.* 50, 2221–2235. doi: 10.1021/ci100335w

Malde, A. K., Zuo, L., Breeze, M., Stroet, M., Poger, D., Nair, P. C., et al. (2011). An Automated force field Topology Builder (ATB) and repository: Version 1.0. *J. Chem. Theory Comput.* 7, 4026–4037. doi: 10.1021/ct200196m

Manz, T. A., and Limas, N. G. (2016). Introducing DDEC6 atomic population analysis: Part 1. Charge partitioning theory and methodology. *RSC Advances* 6, 47771–47801. doi: 10.1039/c6ra04656h

Manz, T. A., and Sholl, D. S. (2010). The Electrostatic Potential in Periodic and Nonperiodic Materials. *J. Chem. Theor. Comput.* 6, 2455–2468.

Manz, T. A., and Sholl, D. S. (2012). Improved atoms-in-molecule charge partitioning functional for simultaneously reproducing the electrostatic potential and chemical states in periodic and nonperiodic materials. *J. Chem. Theory Comput.* 8, 2844–2867. doi: 10.1021/ct3002199

Marenich, A. V., Jerome, S. V., Cramer, C. J., and Truhlar, D. G. (2012). Charge model 5: An extension of hirshfeld population analysis for the accurate description of molecular interactions in gaseous and condensed phases. *J. Chem. Theory Comput.* 8, 527–541. doi: 10.1021/ct200866d

Max, K., and Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *J. Stat. Software* 28, 1–26. doi: 10.1053/j.sodo.2009.03.002

Meinshausen, N. (2006). Quantile Regression Forests. *J. Mach. Learn. Res.* 7, 983–999.

Mobley, D. L. (2013). Experimental and Calculated Small Molecule Hydration Free Energies. *UC Irvine Department* 113 (14), 4533–4537.

Mobley, D. L., and Guthrie, J. P. (2014). FreeSolv: A database of experimental and calculated hydration free energies, with input files. *J. Comput.-Aided Mol. Design* 28, 711–720. doi: 10.1007/s10822-014-9747-x

Mobley, D. L., Bayly, C. I., Cooper, M. D., and Dill, K. A. (2009). Predictions of Hydration Free Energies from All-Atom Molecular Dynamics Simulations †. *J. Phys. Chem. B* 113, 4533–4537. doi: 10.1021/jp806838b

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., et al. (2009). Software News and Updates AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* 30, 2785–2791. doi: 10.1002/jcc

Mortier, W. J., Ghosh, S. K., and Shankar, S. (1986). Electronegativity Equalization Method for the Calculation of Atomic Charges in Molecules. *J. Am. Chem. Soc.* 108, 4315–4320. doi: 10.1021/ja00275a013

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. *J. Cheminf.* 3, 33. doi: 10.1186/1758-2946-3-33

Oostenbrink, C., Villa, A., Mark, A. E., and van Gunsteren, W. F. (2004). A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.* 25, 1656–1676. doi: 10.1002/jcc.20090

Patel, S., and Brooks, C. L. (2004). CHARMM fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations. *J. Comput. Chem.* 25, 1–15. doi: 10.1002/jcc.10355

Patel, S., Mackerell, A. D., and Brooks, C. L. (2004). CHARMM fluctuating charge force field for proteins: II protein/solvent properties from molecular dynamics simulations using a nonadditive electrostatic model. *J. Comput. Chem.* 25, 1504–1514. doi: 10.1002/jcc.20077

Roos, K., Wu, C., Damm, W., Reboul, M., Stevenson, J. M., Lu, C., et al. (2019). OPLS3e: extending force field coverage for drug-like small molecules. *J. Chem. Theory Comput.* 15, 1863–1874. doi: 10.1021/acs.jctc.8b01026

Santos, P. S., Souza, L. K., Araujo, T. S., Medeiros, J. V. R., Nunes, S. C., Carvalho, R. A., et al. (2017). Methylβ-cyclodextrin inclusion complex with $\beta\beta$caryophyllene: Preparation, characterization, and improvement of pharmacological activitie. *ACS Omega* 2, 9080–9094. doi: 10.1021/acsomega.7b01438

Schuler, L. D., Daura, X., and Van Gunsteren, W. F. (2001). An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. *J. Comput. Chem.* 22, 1205–1218. doi: 10.1002/jcc.1078

Schüttelkopf, A. W., and Van Aalten, D. M. F. (2004). PRODRG: A tool for high-throughput crystallography of protein–ligand complexes. *Acta Crystallographica Section D:. Biol. Crystallogr.* 60, 1355–1363. doi: 10.1107/S0907444904011679

Scott, W. R. P., Hünenberger, P. H., Tironi, I. G., Mark, A. E., Billeter, S. R., Fennen, J., et al. (1999). The GROMOS Biomolecular Simulation Program Package. *J. Phys. Chem. A* 103, 3596–3607. doi: 10.1021/jp984217f

Stanke, M., and Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* 33, W465–W467. doi: 10.1093/nar/gki458

Stroet, M., Caron, B., Visscher, K. M., Geerke, D. P., Malde, A. K., Mark, A. E. (2018). Automated Topology Builder Version 3.0: Prediction of Solvation Free Enthalpies in Water and Hexane. *J. Chem. Theory Comput.* 14, 5834–5845. doi: 10.1021/acs.jctc.8b00768

Van Aalten, D. M. (1996). PRODRG, a program for generating molecular topologies and unique molecular descriptors from coordinates of small molecules. *J. Comput.-Aided Mol. Design* 10, 255–262. doi: 10.1007/BF00355047

van Gunsteren, W. F., Billeter, S., Eising, A. A., Hunenberger, P. H., Krüger, P., Mark, A. E., et al. (1996). "*Biomolecular Simulation*," in *The GROMOS96 Manual and User Guide* (Zürich, Switzerland: Vdf Hochschulverlag an der ETH Zürich), 30, 1–1042.

Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004). Development and testing of a general Amber force field. *J. Comput. Chem.* 25, 1157–1174. doi: 10.1002/jcc.20035

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. Graphics Model.* 25, 247–260. doi: 10.1016/j.jmgm.2005.12.005

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

Welch, B. L. (1947). The generalization of "Student's" problem when several different population variances are involved. *Biometrika* 34, 28–35. doi: 10.1093/biomet/34.1-2.28

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.