frontiers in Robotics and AI

PERSPECTIVE
published: 10 January 2020
doi: 10.3389/frobt.2019.00143

Exploring Novel Biologically-Relevant Chemical Space Through Artificial Intelligence: The NCATS ASPIRE Program

*Katharine K. Duncan, Dobrila D. Rudnicki, Christopher P. Austin and Danilo A. Tagle**

National Center for Advancing Translational Sciences, National Institutes of Health, Bethesda, MD, United States

In recent years, artificial intelligence (AI)/machine learning (ML; a subset of AI) have become increasingly important to the biomedical research community. These technologies, coupled to big data and cheminformatics, have tremendous potential to improve the design of novel therapeutics and to provide safe and effective drugs to patients. A National Center for Advancing Translational Sciences (NCATS) program called A Specialized Platform for Innovative Research Exploration (ASPIRE) leverages advances in AI/ML, automated synthetic chemistry, and high-throughput biology, and seeks to enable translation and drug development by catalyzing exploration of biologically active chemical space. Here we discuss the opportunities and challenges surrounding the application of AI/ML to the exploration of novel biologically relevant chemical space as part of ASPIRE.

Keywords: artificial intelligence, machine learning, drug discovery, pharmaceutical development, biomedical research, cheminformatics, translational science

OPEN ACCESS

Edited by:
Akihiro Kishimoto, IBM Research, Ireland

Reviewed by:
Connor Wilson Coley, Massachusetts Institute of Technology, United States
Kunihito Hoki, The University of Electro-Communications, Japan

****Correspondence:***
Danilo A. Tagle
danilo.tagle@nih.gov

Specialty section:
This article was submitted to Computational Intelligence in Robotics, a section of the journal Frontiers in Robotics and AI

Received: *13 May 2019*
Accepted: *09 December 2019*
Published: *10 January 2020*

Citation:
Duncan KK, Rudnicki DD, Austin CP and Tagle DA (2020) Exploring Novel Biologically-Relevant Chemical Space Through Artificial Intelligence: The NCATS ASPIRE Program. Front. Robot. AI 6:143. doi: 10.3389/frobt.2019.00143

INTRODUCTION

Chemical space is incredibly vast; estimates place the number of potential "drug-like" organic molecules between 10^{23} and 10^{60} (Polishchuk et al., 2013). In comparison, the biological space for drug targets is relatively small; the number of protein-coding genes is estimated to be ∼20,000 (Pertea et al., 2018). In order to provide treatments or cures for human diseases, we need to identify novel therapeutics that can modulate the approximately 90% of biological space that is currently undrugged or inaccessible (Barker et al., 2013). However, the current approach to exploring chemical space is extremely limited, requiring manual and labor-intensive synthesis leading to a slow and iterative design-make-test cycles.

Recent innovations in automation and AI/ML create an opportunity for a breakthrough in drug discovery, bringing the goal of a streamlined process for identifying new chemical entities ever closer to becoming reality. To date, scientists have limited ability to predict chemical reactions *a priori* or which molecules will modulate any desired target *ab initio*. Artificial intelligence/machine learning has the potential to allow researchers to uncover areas of chemical space that were previously inaccessible; for example—through the discovery of novel reactions or previously unknown scaffolds. The primary focus of this perspective will be on ways that a recently launched NCATS ASPIRE program seeks to leverage AI/ML to facilitate the exploration of biologically-relevant chemical space. The perspective is not meant to serve as a detailed review of different AI/ML approaches in drug discovery, but to briefly discuss the state of the art in AI/ML,

as well as discuss the challenges to the widespread adoption and use of these new technologies as it applies to the NCATS ASPIRE program.

NCATS ASPIRE PROGRAM: AN OVERVIEW

The National Center for Advancing Translational Sciences recently launched the development of A Specialized Platform for Innovative Research Exploration (ASPIRE) program to capitalize on the recent technological innovations in AI/ML, and potentially disrupt the field of drug discovery (Sittampalam et al., 2018). The NCATS ASPIRE program aims to make the process of exploring chemical space faster, more efficient, and more cost-effective by integrating advances in computer-aided drug design, automated synthetic chemistry, and high-throughput biological screening. This platform will build on the current state of the art to develop innovative algorithms that can predict novel structures capable of modulating specific targets; enable the small-scale synthesis of the suggested molecules; and test these molecules in physiologically relevant biological assays. New data generated through this cycle will then be fed back into the system to help guide the design and synthesis of additional molecules. The NCATS ASPIRE program seeks to move beyond known chemical reactions toward the execution and analysis of novel chemistries. Harnessing advances in chemical laboratory automation, AI/ML, and high-throughput screening, ASPIRE aims to help transform chemistry from an artisanal, empirical practice into a more predictive science. This initiative, in order to be transformational, will require multidisciplinary collaborations among researchers in academic, industrial, and government settings, scientific publishers, funders, and professional societies. NCATS ASPIRE platform and the accompanying tools and technologies, including AI/ML algorithms, chemical laboratory automation, microfluidic flow chemistry, and high throughput screening, will provide a new opportunity to break the translational bottlenecks in chemistry and benefit many areas of science and human health.

The recent launch of National Institutes of Health HEAL (Helping to End Addiction Long-Term) Initiative[SM] provided the opportunity to serve as a pilot for the ASPIRE through prizing competitions via the NCATS ASPIRE Design Challenges[1,2]. Launched in December 2018, these challenges focus on the chemistry and biology of pain, opioid use disorder (OUD), and opioid addiction as a test bed for ASPIRE (**Figure 1**). Based on the input from scientist and other stakeholders, the challenges were design to initially address four major areas of greatest need: (1) Integrated Chemistry Database for Translational Innovation in Pain, OUD and Overdose; (2) Electronic Synthetic Chemistry Portal for Translational Innovation in Pain, OUD and Overdose; (3) Predictive Algorithms for Translational Innovation in Pain, OUD and Overdose; and (4) Biological Assays for Translational Innovation in Pain, OUD and Overdose. In addition, a fifth challenge was conceived for an Integrated Solution where innovators could propose a more comprehensive, single platform solution that combines at least two of the above challenge areas. Readers are directed to 2018 NCATS ASPIRE Design Challenges web page https://ncats.nih.gov/aspire/challenges that contains more detail on how these challenges were structured, and which specific problems they aim to address. It is anticipated that the 2018 NCATS Design Challenges will be followed by a distinct, reduction-to-practice phase in which innovators will build working prototypes from the winning designs of a platform that integrates a chemistry database, electronic synthetic chemistry portal, predictive algorithms, and biological assays. While this initial pilot application for ASPIRE focuses on opioids and the quest for novel treatments for pain, OUD, and overdose, it is anticipated that the developed technologies will be broadly applicable to drug discovery in general and will in the future include additional areas of development such as automated small-scale synthesis. While ambitious, we believe that ASPIRE will enable the next generation of drug developers and medicinal chemists, and produce solutions to previously intractable challenges in medicine.

APPLICATION OF AI TO *DE NOVO* MOLECULAR DESIGN

Researchers have applied AI/ML approaches to many stages of the drug discovery process, from hit identification to lead optimization. Indeed, there has been an explosion in research around the application of these techniques to the problem of chemical synthesis and molecular optimization: a recent review noted 45 papers on molecular optimization in the last 2 years alone (Elton et al., 2019). As stated above, it is beyond the scope of this perspective to recapitulate and discuss the multitude of algorithms and approaches to AI/ML in the literature. Readers are directed toward excellent recent reviews for additional background information on leading AI/ML models, architectures, and techniques as applied to molecular design and optimization (Carpenter et al., 2018; Chen et al., 2018; Engkvist et al., 2018; Lo et al., 2018; Elton et al., 2019; Vamathevan et al., 2019).

One area of focus for researchers is *de novo* molecular design, the computational technique that aims to design novel compounds with a given set of properties such as, for example, high-affinity for a target of interest, or ability to cross the blood-brain barrier. In principle, these *de novo* design algorithms can explore the full span of chemical space and help identify and prioritize molecules that meet a diverse set of criteria (Brown et al., 2019). AI/ML can find patterns and connections within vast data sets, a task too time-consuming for human researchers. AI/ML models continue to improve with larger and robust training sets with which to train and fine-tune the system. The creation and maintenance of robust databases and training sets is therefore crucial to the development of AI/ML models to identify and design new compounds. To date, there are no commonly-accepted training datasets to train models on generating chemical species or for optimizing biological properties (Polykovskiy et al., 2019). Datasets are often pulled

[1]HEAL Initiative: heal.nih.gov
[2]NCATS Challenges: ncats.nih.gov/aspire/challenges/

FIGURE 1 | Graphic representation of the ASPIRE platform workflow. Better therapeutics will be delivered to patients through the effective exploration of chemical space through the application of predictive algorithms, electronic laboratory notebook, integrated database, and biological assays.

from public sources including ChEMBL (Gaulton et al., 2011), ZINC (Sterling and Irwin, 2015), PubChem (Kim et al., 2019), or commercial platforms such as SciFinder[3] and Reaxys[4]. As databases of chemical space grow ever larger, data mining techniques and algorithms to enable chemists to efficiently explore this space are crucial; for example, algorithms to assess synthetic tractability or analyze structural properties would be tremendously valuable (Hoffmann and Gastreich, 2019). Ideal databases for drug discovery include those that are properly curated, complete with positive and negative data, and replete with examples to capture a wide breadth of existing chemical space. When paired with strong design algorithms, such databases can help researchers identify areas of chemical space yet to be explored and quickly sift through what has already been synthesized and tested. Below, we will detail the limitations of existing databases and the difficulties in evaluating the strengths and weaknesses of existing molecular generation algorithms.

Despite the importance of reaction discovery to modern synthetic organic chemistry, less attention has been paid to the application of AI/ML methods to the identification of novel reactions or the optimization of existing reactions. Researchers have demonstrated that AI/ML methods can predict the performance of a synthetic reaction using data from high-throughput experimentation (Ahneman et al., 2018) as well as suitable reaction conditions (Gao et al., 2018). AI/ML methods have been extensively applied to retrosynthesis and computer-aided synthesis planning (Coley et al., 2018; Schwaller et al., 2018; Segler et al., 2018; Baylon et al., 2019). Advantages of such efforts include enabling fully autonomous synthesis and prioritizing of routes with the highest probability of success. Indeed, advances in reaction miniaturization and high throughput experimentation have facilitated the exploration of a larger portion of chemical space (Santanilla et al., 2015). With the capacity to screen ever larger numbers of conditions, researchers need the cheminformatics tools and AI/ML algorithms to make use of the multitude of data. The rate of discovery will further increase when these AI/ML methods are combined with automated experimentation (Häse et al., 2019). Assisting in the optimization and discovery of new reactions, automated platforms will accelerate the synthesis and biological testing of novel compounds (Schneider, 2018). For example, Lilly Research Laboratories recently published a platform called Idea2Data which integrates ML, automated synthesis and high-throughput screening (Nicolaou et al., 2019). Automated synthetic chemistry will play a large role in the further advancement of AI in drug discovery as such closed-loop systems feed an increasing amount of synthetic and biological data back into the system.

CHALLENGES

Despite the many opportunities present in the application of AI/ML to drug discovery, there are several barriers to its widespread acceptance and adoption. These challenges include deficiencies within existing datasets, a lack of interpretability of AI/ML models, potential for models to be self-reinforcing, the need for benchmarks, and the need to increase engagement within the chemistry community.

Need for Improved Synthetic Chemistry and Biological Data Collection/ Dissemination

Central to the development of successful AI/ML models are high-quality data and training sets. Currently, the field of

[3]https://sso.cas.org/as/2bj74/resume/as/authorization.ping
[4]https://www.reaxys.com/#/search/quick

synthetic chemistry is limited by inconsistent use of electronic laboratory notebooks or automated systems that might facilitate the capture of a multitude of additional reaction parameters and conditions. Data concerning failed reactions is often absent from published journal articles and conference presentations. The NCATS ASPIRE program seeks to enable and support the capture and sharing of both positive and negative reaction data, through the development of an electronic synthetic chemistry portal and an integrated chemical database.

Because algorithms require both positive and negative data under a variety of conditions to learn and predict possible reaction outcomes and routes, existing data repositories are severely limited. Widespread adoption and use of electronic laboratory notebooks will improve the quality of databases and training data sets. There is an abundance of chemistry databases, both from commercial and open sources. However, this very abundance of data resources is problematic since there is currently no data standardization, no centralization, and no assurance of quality control. Many available databases are hand-curated, given the current lack of sophisticated machine reading or text extraction capabilities. Coming from multiple different laboratories, the biological activity data is often not comparable or easily correlated with values from other publications (Casciuc et al., 2019). Further, the sparse, heterogeneous nature of these public databases makes it difficult to create single, rigorously defined training sets (Casciuc et al., 2019). Data-reporting standards, including the types of data to be included and specific formats, are needed to facilitate data capture and machine reading, which will in turn help to produce higher-quality datasets to inform ML models. While it is beyond the scope of this perspective to detail what these standards should entail, it is clear that consensus between groups from academia, industry, funders, and publishers is critical for their widespread adoption and acceptance by researchers.

Need for Increased Interpretability and Reliability of Models

The NCATS ASPIRE program seeks to support the development of novel AI/ML algorithms that would aid in the discovery of novel analgesics and treatments for pain, opioid addiction, and overdoses. One challenge to the application of AI algorithms is the lack of interpretability of these models. Many chemists currently perceive these algorithms as a "black box," making it difficult to ascertain how the model arrives at its conclusion(s). Researchers must ensure that AI models derive meaningful conclusions from the data by investigating alternative models to detect potential confounding variables (Chuang and Keiser, 2018).

Need to Address Potentially Self-Reinforcing Nature of MI Models

Further, AI algorithms have the potential to be self-reinforcing: existing systems often prioritize routes based on frequency of appearance in the literature, and not necessarily because they are the best possible routes (Jordan, 2018). As models suggest these common transformations and are increasingly used by chemists, they become more likely to be suggested by the same AI systems. This feedback loop could then lead to an overreliance on certain reactions and reduce synthetic creativity.

Need for Benchmarks

With the recent proliferation of papers and approaches describing the application of deep learning techniques to drug discovery, the challenge becomes how best to evaluate and compare them and to discern their relative strengths and weaknesses. A series of standard metrics and benchmarks that would facilitate the evaluation and comparison of new and existing models was proposed recently (Wu et al., 2018; Brown et al., 2019; Polykovskiy et al., 2019). Suggested metrics include fragment similarity, scaffold similarity, nearest neighbor similarity, internal diversity, and Frechnet ChemNet Distance (Polykovskiy et al., 2019). Researchers have also proposed dividing benchmarks into distribution-learning benchmarks, including validity and novelty, and goal-directed benchmarks (Brown et al., 2019). It is also important to note that benchmarks for molecular optimization have different demands and these need to be addressed. AI researchers should strive to make their codes and datasets widely accessible to the larger community to facilitate comparison and evaluation by others. These proposed benchmarking standards and open sharing will provide the framework and means to evaluate the quality and diversity of the molecules generated by the models. While the current NCATS ASPIRE program does not directly address the need for benchmarking, we recognize its importance for the success of the program and the field.

Need to Encourage Engagement by Chemists

Some scientists are skeptical that a machine can learn and execute on the nuances of medicinal and organic chemistries and believe that the power of AI/ML is overestimated (Lemonick, 2018). As with any new technology, this skepticism naturally leads to a reluctance to engage with or adopt AI models. Even if AI is "overhyped," as most respondents to a recent C&EN survey indicated (Lemonick, 2018), it does offer some opportunities and advantages that warrant attention from synthetic and medicinal chemists. Increasing engagement with and use of AI/ML by chemists depends on increased education around these advantages and opportunities, particularly with regard to how these new technologies might make enhance the day to day activities of chemists. *De novo* molecule generation and synthetic route planning will assist chemists in prioritizing analogs to synthesize and thus shorten the lead optimization process. Further, these innovations will provide the basis for AI-enabled automated laboratories (Sanchez-Lengeling and Aspuru-Guzik, 2018). Automated chemistry laboratories would liberate chemists from much of the mundane, physical burden of weighing out reagents, setting up reactions, and purifying final products. With human error minimized, chemical reactions should be more reproducible and efficient. Untethered to the bench, chemists would be available to spend time on more difficult synthetic challenges or novel intellectual pursuits. Automated laboratories would also make areas of research more accessible to individuals

with disabilities (Lemonick, 2019). Increased data sharing, as discussed above, would reduce the duplication of efforts and facilitate the adoption of novel methods and reactions in the laboratory. Together, these benefits would transform the practice of medicinal chemistry and reduce the time and cost of bringing new chemical probes, to sciences and new therapeutics to patients.

CONCLUSIONS AND FUTURE OUTLOOK

The application of AI to the design of novel molecules and reaction conditions has the power to transform many aspects of drug discovery, including the identification of new biological targets, novel scaffolds, and improved synthetic routes. The NCATS ASPIRE program will utilize the power of emerging technologies, including AI, recent innovations in automated synthesis, liquid handling and microfluidics, and high-throughput screening to effectively assay chemical space and identify novel biologically active small molecules. By accelerating the design-make-test cycle and delivering novel molecules more efficiently, ASPIRE has the potential to reduce or eliminate the bottlenecks in chemical biology and pharmaceutical development. The ASPIRE Design Challenges are an important first step in implementing the ASPIRE program, addressing the current need for improved chemical databases, electronic laboratory notebooks, biological assays, and algorithms. The implementation of ASPIRE through prizing competition allows for future expansion of the program in response to the need for new tools and technology during the program's course, as well as newly identified scientific roadblock that may need to be addressed in the future to achieve the goals of the ASPIRE. Some of the components that we currently see as being important to be included in the program and assure its success include data standardization, consensus on descriptors and metadata needed to enable automated synthetic technologies, and improved laboratory automation equipment with user-friendly interfaces. The tools and technologies developed as part of the NCATS ASPIRE Design Challenges will initially be focused on analgesics with minimal addictive properties, as part of the NIH HEAL Initiative[SM]. However, we anticipate that the developed and eventually widely disseminated tools and technologies will be transferable to other diseases and disorders in the future.

Further technological advances in AI will require a concerted effort across the chemistry community, involving funders, academic institutions, and industry to bring the field fully into the twenty first century. From software to hardware developers to the next generation of bench scientists, the field needs to collaborate on the best ways to capture and utilize the data of chemical reactions and facilitate the identification of new chemistries and the discovery of novel therapeutics. These "labs of the future" will harness the power of automated synthesis platforms, AI, and high-throughput biological assays to increase the speed and safety, while reducing the time and cost, required to bring new therapies to patients. Once widely available and utilized, these platforms will help make the dream of precision medicine, medications tailored to a particular patient with a particular disease state, a reality.

AUTHOR CONTRIBUTIONS

KD wrote and edited the manuscript. DR and DT conceived the extramural ASPIRE concept and, together with CA, edited the manuscript.

ACKNOWLEDGMENTS

The authors wish to thank Kyle Brimacombe for the design of the ASPIRE Challenges badge that is part of **Figure 1**.

REFERENCES

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* 360:186190. doi: 10.1126/science.aar5169

Barker, A., Kettle, J. G., Nowak, T., and Pease, J. E. (2013). Expanding medicinal chemistry space. *Drug Discov. Today* 18, 298–304. doi: 10.1016/j.drudis.2012.10.008

Baylon, J. L., Cilfone, N. A., Gulcher, J. R., and Chittenden, T. W. (2019). Enhancing retrosynthetic reaction prediction with deep learning using multiscale reaction classification. *J. Chem. Inf. Model.* 59, 673–688. doi: 10.1021/acs.jcim.8b00801

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

Carpenter, K. A., Cohen, D. S., Jarrell, J. T., and Huang, X. (2018). Deep learning and virtual drug screening. *Fut. Med. Chem.* 10:21. doi: 10.4155/fmc-2018-0314

Casciuc, I., Horvath, D., Gryniukova, A., Tolmachova, K. A., Vasylchenko, A. V., Borysko, P., et al. (2019). Pros and cons of virtual screening based on public "Big Data": *in silico* mining for new bromodomain inhibitors. *Eur. J. Med. Chem.* 165, 258–272. doi: 10.1016/j.ejmech.2019.01.010

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

Chuang, K. V., and Keiser, M. J. (2018). Adversarial controls for scientific machine learning. *ACS Chem. Biol.* 13, 2819–2821. doi: 10.1021/acschembio.8b00881

Coley, C. W., Green, W. H., and Jensen, K. F. (2018). Machine learning in computer-aided synthesis planning. *Acc. Chem. Res.* 51, 1281–1289. doi: 10.1021/acs.accounts.8b00087

Elton, D. C., Boukouvalas, Z., Fuge, M. D., and Chung, P. W. (2019). Deep learning for molecular generation and optimization – a review of the state of the art. *arXiv*. doi: 10.1039/C9ME00039A

Engkvist, O., Norrby, P.-O., Selmi, N., Lam, Y.-H., Peng, Z., Sherer, E. C., et al. (2018). Computational prediction of chemical reactions: current status and outlook. *Drug Discov. Today* 23, 1203–1218. doi: 10.1016/j.drudis.2018.02.014

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 Cent. Sci.* 4, 1465–1476. doi: 10.1021/acscentsci.8b00357

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., et al. (2011). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40, D1100–D1107. doi: 10.1093/nar/gkr777

Häse, F., Roch, L. M., and Aspuru-Guzik, A. (2019). Next-generation experimentation with self-driving laboratories. *Trends Chem.* 1, 282–291. doi: 10.1016/j.trechm.2019.02.007

Hoffmann, T., and Gastreich, M. (2019). The next level in chemical space navigation: going far beyond enumerable compound libraries. *Drug Discov. Today.* 24, 1148–1156. doi: 10.1016/j.drudis.2019.02.013

Jordan, A. (2018). Artificial intelligence in drug design—the storm before the calm? *ACS Med. Chem. Lett.* 9, 1150–1152. doi: 10.1021/acsmedchemlett.8b00500

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., et al. (2019). PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* 47, D1102–D1109. doi: 10.1093/nar/gky1033

Lemonick, S. (2018). *Is Machine Learning Overhyped: CHEMISTS Weigh in on the Technique's Possibilities and Its Pitfalls*. Chemical and Engineering News, 96. Available online at: https://cen.acs.org/physical-chemistry/computational-chemistry/machine-learning-overhyped/96/i34 (accessed April 18, 2019).

Lemonick, S. (2019). *Artificial Intelligence Tools Could Benefit Chemists With Disabilities. So Why Aren't They*? Chemical and Engineering News, 97. Available online at: https://cen.acs.org/careers/diversity/Artificial-intelligence-tools-benefit-chemists/97/i11 (accessed April 10, 2019).

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

Nicolaou, C. A., Humblet, C., Hu, H., Martin, E. M., Dorsey, F. C., Castle, T. M., et al. (2019). Idea2Data: toward a new paradigm for drug discovery. *ACS Med. Chem. Lett.* 10, 278–286. doi: 10.1021/acsmedchemlett.8b00488

Pertea, M., Shumate, A., Pertea, G., Varabyou, A., Breitwieser, F. P., Chang, Y. C., et al. (2018). CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol.* 19:208. doi: 10.1186/s13059-018-1590-2

Polishchuk, P. G., Madzhidov, T. I., and Varnek, A. (2013). Estimation of the size of drug-like chemical space based on GDB-17 data. *J. Comput. Aided Mol. Des.* 27, 675–679. doi: 10.1007/s10822-013-9672-4.

Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev, S., et al., (2019). Molecular sets (MOSES): a benchmarking platform for molecular generation Models. *arXiv*.

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

Santanilla, A. B., Regalado, E. L., Pereira, T., Shevlin, M., Bateman, K., Campeau, L.-C., et al. (2015). Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science* 347, 49–53. doi: 10.1126/science.1259203

Schneider, G. (2018). Automating drug discovery. *Nat. Rev. Drug Discov.* 17, 97–113. doi: 10.1038/nrd.2017.232

Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Bekas, C., and Lee, A. A. (2018). Molecular transformer for chemical reaction prediction and uncertainty estimation. *ChemRXiv*. doi: 10.26434/chemrxiv.7297379.v1

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

Sittampalam, G. S., Rudnicki, D., Tagle, D. A., Simeonov, A., and Austin, C. P. (2018). Mapping biologically active chemical space to accelerate drug discovery. *Nat. Rev. Drug Discov.* 18, 83–84. doi: 10.1038/d41573-018-00007-2

Sterling, T., and Irwin, J. J. (2015). ZINC 15 – ligand discovery for everyone. *J. Chem. Inf. Model.* 55:2324. doi: 10.1021/acs.jcim.5b00559

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., et al. (2019). Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* 18, 463–477. doi: 10.1038/s41573-019-0024-5

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., et al. (2018). Moleculenet: a benchmark for molecular machine learning. *Chem. Sci.* 9, 513–530. doi: 10.1039/C7SC02664A

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.