EDITORIAL: Chemical Compound Space Exploration by Multiscale High-Throughput Screening and Machine Learning

According to various estimates, the number of possible chemical compounds is in the range of 10^{18}–10^{200}. A popular estimate of 10^{60} refers to molecules composed of C, H, N, O, and H atoms, containing no more than 4 rings and weighting less than 500 Da.[1] Several other "theoretical" subspaces populated by small organic (often, drug-like) molecules have been enumerated, e.g., a more conservative estimate of 3.4×10^9 for molecules with ≤100 carbon atoms[2] and the "Chemical Universe Database" GDB-17 with 166.4 billion molecules with up to 17 C, N, O, S, and halogen atoms.[3,4] Across the entire chemical space, ca. 219 million organic substances, alloys, coordination compounds, minerals, mixtures, polymers, and salts have been published and are recorded in the Chemical Abstracts Service (CAS) registry.[5]

Although seemingly a huge number, 219 million is but a speck of dust in the practically infinite chemical universe. Such endless possibilities come with a burden of choice: which molecule or material is the best (most efficient, cheapest, most sustainable, ...) for a given practical use? Chemists have come up with diverse solutions to this problem over the centuries, but a profound shift away from building upon prior experience (for example, introducing new substituents into a known catalyst) and toward a less biased and more broad chemical space exploration has only come about comparatively recently. The immense growth in computing power and memory, the variety and availability of theoretical methods and their implementations, new chemical synthesis approaches and laboratory automation, and stellar advances in artificial intelligence are all factors contributing to this shift. Today, combinatorial structure generation and property computation using automated multiscale workflows enable high-throughput screening (HTS) of millions of compounds at the cost we used to pay for computing a single chemically accurate energy profile for a reaction between relatively small compounds only some 20 years ago. Moreover, generative and predictive machine learning (ML) models enable targeted inverse molecular design and allow estimation of chemical properties across even larger regions of the chemical universe.[6−11]

The latest advances in AI-driven molecular science were central to two recent meetings in the historic city of Heidelberg in the south of Germany: the second SIMPLAIX Workshop on Machine Learning for Multiscale Molecular Modeling (https://simplaix-workshop2024.h-its.org/) and the Chemical Compound Space Conference 2024 (CCSC2024, https://ccsc2024.github.io/, Figure 1). A broad range of scientific themes, from developing new machine learning architectures for studying molecular properties to ML applications in biomolecular simulations and materials discovery, was covered. However, exploration of chemical space served as a leitmotif for many of the lectures, posters, and discussions. Despite the meteoric pace of progress in the

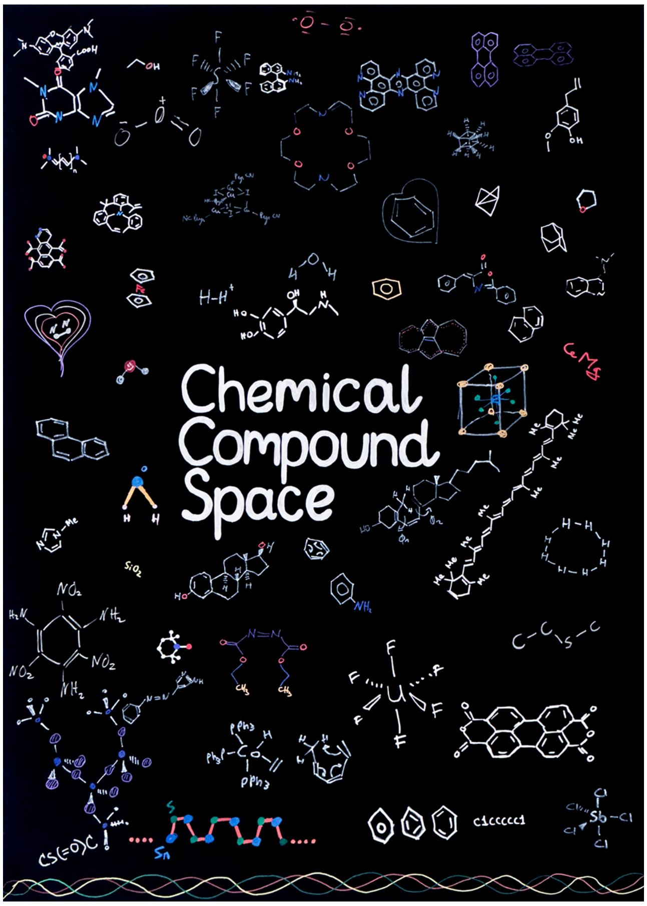

Figure 1. Illustration of chemical compound space created by the attendees during the CCSC2024 conference. Photo credit: John M. Lindner.

field of chemical big data and machine learning, many experts agree that efficient, comprehensive, and unbiased exploration of this immense space remains elusive. The very fast pace of methodological developments in this area was named among the key obstacles to chemical space exploration, with parallels being drawn between the "alphabet soup" of density functionals (a term coined by Kieron Burke in 2007) and the multitude of new ML potentials, architectures, and representations published today. Only time will show whether the field will eventually converge on a handful of popular models. The "garbage in, garbage out" problem is equally central to chemical space exploration, as the reliability of predictions for new systems is a direct outcome of the quality of the training data. Consequently, there is a pressing need for automated approaches to benchmarking, comparing, and quantifying the uncertainties of ML models in chemistry. Finally, even if and when promising new molecules and materials are discovered *in silico*, their stability and synthetic accessibility, challenging to predict for truly novel systems, become key to successful experimental validation. Mining the literature and training ML models on experimental data offer a potential solution to this issue, yet both these approaches are hindered by barriers to the availability and accessibility of such data and the lack of FAIR data standards when reporting and structuring it.

In the light of these challenges, the *Journal of Chemical Information and Modeling* (JCIM) invites authors to submit contributions to a Virtual Special Issue (VSI) on the topic of **"Chemical Compound Space Exploration by Multiscale High-Throughput Screening and Machine Learning"**. This VSI recognizes the vertiginous developments in the field during the last three years since the JCIM special issue on Reaction Informatics and Chemical Space.[12] All manuscript types published by JCIM, including articles, perspectives, viewpoints, reviews, letters, and application notes, are welcome. For more information on manuscript types and how to submit, please visit the journal's Web site.

Submissions will be received through January 31, 2025. All articles submitted under this VSI will be peer-reviewed to ensure they fit the scope of the Virtual Special Issue and that they meet the high scientific publishing standards of the *Journal of Chemical Information and Modeling* (more information can be found in previous editorials[13,14]). If accepted, publications will go online as soon as possible and be published in the next available issue. Publications on this topic will be gathered into a Virtual Special Issue and widely promoted thereafter.

Ganna Gryn'ova orcid.org/0000-0003-4229-939X
Tristan Bereau orcid.org/0000-0001-9945-1271
Carolin Müller orcid.org/0000-0002-5968-2216
Pascal Friederich orcid.org/0000-0003-4465-1465
Rebecca C. Wade orcid.org/0000-0001-5951-8670
Ariane Nunes-Alves
Thereza A. Soares orcid.org/0000-0002-5891-6906
Kenneth Merz, Jr.

AUTHOR INFORMATION

Notes

Views expressed in this editorial are those of the authors and not necessarily the views of the ACS.

ACKNOWLEDGMENTS

G.G. gratefully acknowledges funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement no. 101042290 PATTERNCHEM). T.B. acknowledges support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2181/1-390900948 (the Heidelberg STRUCTURES Excellence Cluster). G.G, T.B., P.F., and R.C.W. gratefully acknowledge the financial support of the Klaus Tschira Stiftung gGmbH for SIMPLAIX. T.A.S. acknowledges funding from FAPESP (Grant 2022/07231-7), CNPq (Productivity fellowship), and the RCN (Grant 26269). A.N.A. acknowledges funding from DFG under Germany's Excellence Strategy EXC 2008/1-390540038 - UniSysCat.

REFERENCES

(1) Kirkpatrick, P.; Ellis, C. Chemical space. *Nature* **2004**, *432*, 823.

(2) Drew, K. L. M.; Baiman, H.; Khwaounjoo, P.; Yu, B.; Reynisson, J. Size estimation of chemical space: how big is it? *J. Pharm. Pharmacol.* **2012**, *64*, 490−495.

(3) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864−2875.

(4) Warr, W. A.; Nicklaus, M. C.; Nicolaou, C. A.; Rarey, M. Exploration of ultralarge compound collections for drug discovery. *J. Chem. Inf. Model.* **2022**, *62*, 2021−2034.

(5) https://www.cas.org/cas-data/cas-registry, accessed on 2024-07-16.

(6) von Lilienfeld, O.; Müller, K.; Tkatchenko, A. Exploring chemical compound space with quantum-based machine learning. *Nat. Rev. Chem.* **2020**, *4*, 347−358.

(7) Pollice, R.; dos Passos Gomes, G.; Aldeghi, M.; Hickman, R. J.; Krenn, M.; Lavigne, C.; Lindner-D'Addario, M. i.; Nigam, A.; Ser, C. T.; Yao, Z.; Aspuru-Guzik, A. Data-driven strategies for accelerated materials design. *Acc. Chem. Res.* **2021**, *54*, 849−860.

(8) Nandy, A.; Duan, C.; Taylor, M. G.; Liu, F.; Steeves, A. H.; Kulik, H. J. Computational discovery of transition-metal complexes: From high-throughput screening to machine learning. *Chem. Rev.* **2021**, *121*, 9927−10000.

(9) Rosen, A. S.; Iyer, S. M.; Ray, D.; Yao, Z.; Aspuru-Guzik, A.; Gagliardi, L.; Notestein, J. M.; Snurr, R. Q. Machine learning the quantum-chemical properties of metal−organic frameworks for accelerated materials discovery. *Matter* **2021**, *4*, 1578−1597.

(10) Bilodeau, C.; Jin, W.; Jaakkola, T.; Barzilay, R.; Jensen, K. F. Generative models for molecular discovery: Recent advances and challenges. *WIREs Comput. Mol. Sci.* **2022**, *12*, No. e1608.

(11) Anstine, D. M.; Isayev, O. Generative models as an emerging paradigm in the chemical sciences. *J. Am. Chem. Soc.* **2023**, *145*, 8736−8750.

(12) Rarey, M.; Nicklaus, M. C.; Warr, W. Special issue on reaction informatics and chemical space. *J. Chem. Inf. Model.* **2022**, *62*, 2009−2010.

(13) Wei, G.-W.; Zhu, F.; Merz, K. M. Editorial on machine learning. *J. Chem. Inf. Model.* **2022**, *62*, 3941−3941.

(14) Merz, K. M. J.; Amaro, R.; Cournia, Z.; Rarey, M.; Soares, T.; Tropsha, A.; Wahab, H. A.; Wang, R. Editorial: Method and data sharing and reproducibility of scientific results. *J. Chem. Inf. Model.* **2020**, *60*, 5868−5869.