PNAS SPECIAL FEATURE: INTRODUCTION OPEN ACCESS

Machine learning in chemistry

Pablo G. Debenedetti[a,1], Juan J. de Pablo[b,c], and George C. Schatz[d]

Machine learning (ML), a subfield of artificial intelligence (AI), involves the development of algorithms that enable computer systems to learn from data and perform specific tasks or make predictions without being explicitly programmed for such tasks. ML has a broad range of applications in chemistry, including protein design (1), drug and materials discovery (2), property prediction (3), acceleration of quantum-accurate simulations (4), computational catalysis and reaction engineering (5), design of synthetic pathways and processes (6), and automation of complex spectral assignments (7).

The seven research articles and two Perspectives included in this Special Feature on Machine Learning in Chemistry (8–16) are illustrative of the transformative influence of data-driven approaches on contemporary chemical research. The two Perspectives cover generative AI in computational chemistry (8) and equivariant neural networks in chemistry and physics (9). Within the broad categories of materials design and property prediction, topics covered in the research articles address protein design (10), transition metal complex design (11), quantum-accurate, data-driven modeling of calcium carbonate in solution and in the solid state (12), and mechanisms of thermal transport in crystalline inorganic perovskites (13). Methodological learning advances presented include compact vectorized representation of chemical environments leading to reduced model training and prediction compute times (14), ensuring viable synthetic pathways in model-generated molecules (15), and enforcing spatial distance constraints between atomic nuclei in model-generated molecules (16). In what follows, we summarize each contribution.

The Perspective by Tiwary et al. (8) offers a comprehensive overview of generative AI methods in computational chemistry. Approaches that generate new outputs (e.g., inferring phase transitions) by learning from existing data (e.g., limited configurational observations) are referred to as generative AI methods. The authors review fundamental concepts and definitions in generative AI and computational chemistry. They then provide an overview of generative AI methods, including autoencoders, adversarial networks, reinforcement learning, flow-based methods, and large language models. Selected applications in computational quantum chemistry, structural biology, and biophysics are discussed. Finally, the authors address desirable characteristics for generative AI methods in chemistry, emphasizing the ability to predict emergent chemical phenomena as an important objective.

The Perspective by Kondor (9) provides an overview of the mathematical foundations and practical construction of equivariant neural networks (ENNs) for applications in physics and chemistry. The traditional AI models used in general domains, such as language or image recognition, do not explicitly incorporate physical symmetries. In physics and chemistry, symmetries such as translational, rotational, and identical particle exchange are exact and critical for models to make physically meaningful and generalizable predictions. ENNs are models designed to inherently respect such symmetries. They are built using group representation theory, enabling them to transform inputs and outputs in a mathematically consistent way under group actions. The fundamental concepts discussed in this Perspective include invariance and equivariance: Invariant models output the same result regardless of transformations, whereas equivariant models output results that transform predictably under transformations (e.g., forces rotate consistently with atomic positions). The review also discusses group and irreducible representations, including the fundamental building blocks used to decompose and construct ENNs. As clearly outlined in this Perspective, ENNs have become central to the representation of molecular systems with translation, rotation, and permutation symmetry and have established their usefulness in force field learning and property prediction in computational chemistry and physics.

Sevgen et al. (10) tackle a fundamental problem in protein engineering, namely the discovery of sequences with desired functionality (the sequence-function problem). Combining two generative modeling approaches, namely transformer-based protein language models and variational autoencoders, they introduce the Protein Transformer Variational AutoEncoder model for data-driven protein design. Testing the model's designs experimentally, the authors discover a phenylalanine hydroxylase (PAH) enzyme mutant with 2.5 times the catalytic activity relative to the human PAH wild type, and a γ-carbonic anhydrase (γ-CA) enzyme with a 61 °C melting temperature elevation relative to the highest similarity natural γ-CA, with stability at industrially relevant conditions for enzymatic carbon capture. The approach can be applied generically to other machine learning–guided directed evolution efforts (17) and enables direct learning of the sequence-to-function mapping, in the absence of structure data.

In their paper, Toney et al. aim to generate three-dimensional (3D) structures of transition metal complexes (TMCs) with predicted metal-ligand coordination (11). They use a large dataset of ligands of known coordination from

Author affiliations: [a]Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ 08544; [b]Department of Chemical and Biomolecular Engineering, Tandon School of Engineering, New York University, Brooklyn, NY 11201; [c]Department of Computer Science, Courant Institute of Mathematical Sciences, New York University, New York, NY 10012; and [d]Department of Chemistry, Northwestern University, Evanston, IL 60208

Author contributions: P.G.D., J.J.d.P., and G.C.S. designed research; and wrote the paper.

The authors declare no competing interest.

P.G.D., J.J.d.P. and G.C.S. are organizers of this Special Feature.

[1]To whom correspondence may be addressed. Email: pdebene@princeton.edu.

Published October 6, 2025.

experimental structures of TMCs in the Cambridge Structural Database (CSD) to train and validate a graph neural network, with the goal of predicting the number and identities of ligand coordinating atoms in these complexes. With extensive curation, the neural network is used to predict ligand-metal coordination for previously unknown complexes, and these are validated by comparison with density functional theory (DFT) calculations. A Simplified Molecular Input Line Entry System (SMILES) representation of the complexes is used to define molecular structure for the training, while CSD data are based on atomic positions in 3D, so an important component of work is to assess the ability of the neural network to connect SMILES to 3D. A careful accuracy analysis is performed based on the ability to reproduce the total number and individual identities of ligand-coordinating atoms. The accuracy is found to vary with the number of ligands and the choice of ligand and metal.

Calcium carbonate is key to carbon sequestration technology, the regulation of ocean acidity, and biomineralization. Piaggi et al. (12) develop a first-principles machine learning model to study the formation of calcium carbonate from aqueous solution using molecular dynamics simulation. The model strongly constrained and appropriately normed-ML (SCAN-ML) is trained on ab initio DFT forces and energies within the SCAN approximation for the exchange and correlation functional (18). The approach naturally allows for the occurrence of chemical reactions, which are essential in the case of calcium carbonate formation. SCAN-ML captures a broad range of structural and dynamic properties of single ions in solution and calcium carbonate solid phases with an accuracy that surpasses state-of-the-art force fields and compares very well with experiments, while also capturing ion pairing free energy curves and the structure of the calcite–water interface.

The low-temperature thermal conductivities of crystals and glasses exhibit distinct temperature dependencies (e.g., $\sim T^3$ for crystals, $\sim T^2$ for glasses). However, some crystalline inorganic perovskites exhibit glassy thermal conductivities at low temperatures. The origin of this behavior is not well understood. Zeng et al. (13) study the thermal conductivity of the crystalline perovskite $Cs_3Bi_2I_6Cl_3$, using path integral molecular dynamics in conjunction with machine learning potentials. The authors are able to reproduce experimentally observed trends. They find that the system exhibits pronounced lattice distortions at low temperatures, which, the authors suggest, may be due to large atomic size mismatch.

In developing many-body potential energy functions for molecules and solids, there is a tradeoff between how much physics is incorporated into the function used to represent the potential, usually expressed as the kernel that connects points in configuration space to energies used in training, and how much data are needed for generating a meaningful potential. The paper by Khan and von Lilienfeld (14) provides important new results related to improving the physics side of the story through the development of generalized convolutional many-body distribution functionals (cMBDF) as compute- and data-efficient atomic representations of the kernels. In this work, a kernel ridge regression (KRR)-based machine learning approach is used to represent atomic densities weighted by interaction potentials. Representing densities in terms of Gaussians greatly simplifies the analytical representation of the functionals, so that many results can be preevaluated on a grid and stored for later use. cMBDF is found to require 32 times less training data than atom-centered symmetry functions (19) for the same accuracy, while still being faster to evaluate.

The paper by Gao et al. (15) introduces a novel generative AI framework designed to efficiently generate synthesizable molecules by explicitly designing their synthetic pathways, overcoming the key limitation of prior models that often propose molecules that are impossible to synthesize. The key advance in this paper is to combine transformer architectures with denoising diffusion models to generate synthetic routes, rather than just molecular structures. In doing so, the proposed approach ensures synthetic feasibility by operating within a chemical space defined by purchasable building blocks and reliable reaction templates; it addresses a critical bottleneck in generative molecular design by ensuring that every generated molecule is synthetically tractable, bridging AI design outputs and experimental feasibility.

The paper by Liu et al. (16) introduces a new generative AI model for structure-based drug design. It aims to overcome a key limitation in current models: The generation of molecules that contain atomic overlaps (e.g., atoms placed unrealistically close, violating physical constraints). Some existing models of molecules treat atoms as solid points without considering their electron cloud sizes, leading to unrealistic structures where atoms collide. The proposed model, NucleusDiff, consists of a diffusion-based generative model that explicitly incorporates the atomic nuclei position and their corresponding van der Waals radii by using discretized mesh points. The resulting improvements are demonstrated in the context of several applications, including improved binding affinity by up to 22% in general benchmarks and 21% for COVID-19 targets. The proposed approach is phenomenological in nature and relies on discretized mesh points to approximate continuous manifolds; the authors explain how future improvements could seek to integrate first-principles quantum mechanical representations.

We hope that the nine contributions that make up this Special Feature are able to convey the accomplishments, possibilities, and challenges that define the exciting scientific frontier at the confluence of chemistry and machine learning.

1. A. Madani *et al.*, Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* **41**, 1099–1106 (2023).
2. O. A. von Lilienfeld, K.-R. Müller, A. Tkatchenko, Exploring chemical compound space with quantum-based machine learning. *Nat. Rev. Chem.* **4**, 347–358 (2020).
3. S. Boobier, D. R. J. Hose, A. J. Blacker, B. N. Nguyen, Machine learning with physicochemical predictions: Solubility prediction in organic solvents and water. *Nat. Commun.* **11**, 5753 (2020).
4. P. M. Piaggi, J. Weis, A. Z. Panagiotopoulos, P. G. Debenedetti, R. Car, Homogeneous ice nucleation in an ab-initio machine-learning model of water. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2207294119 (2022).
5. R. Tran *et al.*, The Open Catalyst 2022 (OC22) dataset and challenges for oxide electrocatalysts. *ACS Catal.* **13**, 3066–3084 (2023).
6. M. A. Al Massud, A. Aria, Y. Wang, J. Hu, Y. Tian, Machine learning-aided design using limited experimental data: A microwave-assisted ammonia synthesis case study. *Amer. Inst. Chem. Eng. J.* **71**, e18621 (2024).
7. P. Klukowski, R. Riek, P. Güntert, Rapid protein assignments and structures from raw NMR spectra with the deep learning technique ARTINA. *Nat. Commun.* **13**, 6151 (2022).
8. P. Tiwary *et al.*, Generative artificial intelligence for computational chemistry: A roadmap to predicting emergent phenomena. *Proc. Natl. Acad. Sci. U.S.A.***122**, e2415655121 (2025).
9. R. Kondor, The principles behind equivariant neural networks for physics and chemistry. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2415656122 (2025).

10. E. Sevgen *et al.*, ProT-VAE: Protein transformer variational autoencoder for functional protein design. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2408737122 (2025).
11. J. W. Toney *et al.*, Graph neural networks for predicting metal–ligand coordination for transition metal complexes. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2415658122 (2025).
12. P. M. Piaggi, J. D. Gale, P. Raiteri, Ab initio machine learning simulation of calcium carbonate from aqueous solutions to the solid state. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2415663122 (2025).
13. Z. Zeng *et al.*, Lattice distortion leads to glassy thermal transport in crystalline $Cs_3Bi_2I_6Cl_3$. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2415664122 (2025).
14. D. Khan, O. A. von Lilienfeld, Generalized convolutional many body distribution functional representations. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2415662122 (2025).
15. W. Gao, S. Luo, C. W. Coley, Generative artificial intelligence for navigating synthesizable chemical space. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2415665122 (2025).
16. S. Liu *et al.*, Manifold-constrained nucleus-level denoising diffusion model for structure-based drug design. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2415666122 (2025).
17. K. K. Yang, Z. Wu, F. H. Arnold, Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **16**, 687–694 (2019).
18. J. Sun, A. Ruzsinszky, J. P. Perdew, Strongly constrained and appropriately normed semilocal density functional. *Phys. Rev. Lett.* **115**, 036402 (2015).
19. J. Behler, Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **134**, 074106 (2011).