Chem Soc Rev

REVIEW ARTICLE

View Article Online
View Journal

Cite this: DOI: 10.1039/d5cs00053j

Accelerating battery innovation: AI-powered molecular discovery

Yu-Chen Gao, [a] Xiang Chen, *[abc] Yu-Hang Yuan, [a] Yao-Peng Chen, [a] Yi-Lin Niu, [a] Nan Yao, [a] Yan-Bin Gao, [a] Wei-Lin Li [a] and Qiang Zhang *[abc]

Received 5th May 2025

DOI: 10.1039/d5cs00053j

rsc.li/chem-soc-rev

The global energy transition urgently demands advanced battery technologies to address current climate challenges, where molecular engineering plays a pivotal role in optimizing performance metrics such as energy density, cycling lifespan, and safety. This review systematically examines the integration of artificial intelligence (AI) into molecular discovery for next-generation battery systems, addressing both transformative potential and sustainability challenges. Firstly, multidimensional strategies for molecular representation are delineated to establish machine-readable inputs, serving as a prerequisite for AI-driven molecular discovery (Section 2). Subsequently, AI algorithms are systematically summarized, encompassing classical machine learning, deep learning, and the emerging class of large language models (Section 3). Next, the substantial potential of AI-powered predictions for key electrochemical properties is illustrated, including redox potential, viscosity, and dielectric constant (Section 4). Through paradigmatic case studies, significant applications of AI in molecular design are elucidated, spanning chemical knowledge discovery, high-throughput virtual screening, oriented molecular generation, and high-throughput experimentation (Section 5). Finally, a general conclusion and a critical perspective on current challenges and future directions are presented, emphasizing the integration of molecular databases, algorithms, computational power, and autonomous experimental platforms. AI is expected to accelerate molecular design, thereby facilitating the development of next-generation battery systems and enabling sustainable energy innovations.

1. Introduction

1.1. Batteries and molecular discovery

The escalating impacts of climate change have elevated the global energy transition to an existential priority, given that energy systems account for 75% of anthropogenic greenhouse gas emissions.[1] To meet the 1.5 °C target of the Paris

[a] Beijing Key Laboratory of Complex Solid State Batteries & Tsinghua Center for Green Chemical Engineering Electrification, Department of Chemical Engineering, Tsinghua University, Beijing 100084, P.R. China.
E-mail: xiangchen@mail.tsinghua.edu.cn, zhang-qiang@mails.tsinghua.edu.cn
[b] Institute for Carbon Neutrality, Tsinghua University, Beijing 100084, P.R. China
[c] Innovation Center for Smart Solid-State Batteries, Yibin, P.R. China

Yu-Chen Gao

Yu-Chen Gao received his bachelor's degree from Tianjin University in 2022. He is currently a PhD student in the Department of Chemical Engineering at Tsinghua University. His research focuses on using data-driven methods to understand the chemical mechanism in rechargeable batteries and artificial intelligence to accelerate molecular design and material discovery.

Xiang Chen

Xiang Chen received his bachelor's and PhD degrees from the Department of Chemical Engineering at Tsinghua University in 2016 and 2021, respectively. He is currently an associate professor at Tsinghua University. His research interests focus on understanding the chemical mechanisms and materials science in rechargeable batteries through multi-scale simulation, artificial intelligence, and autonomous experimentation.

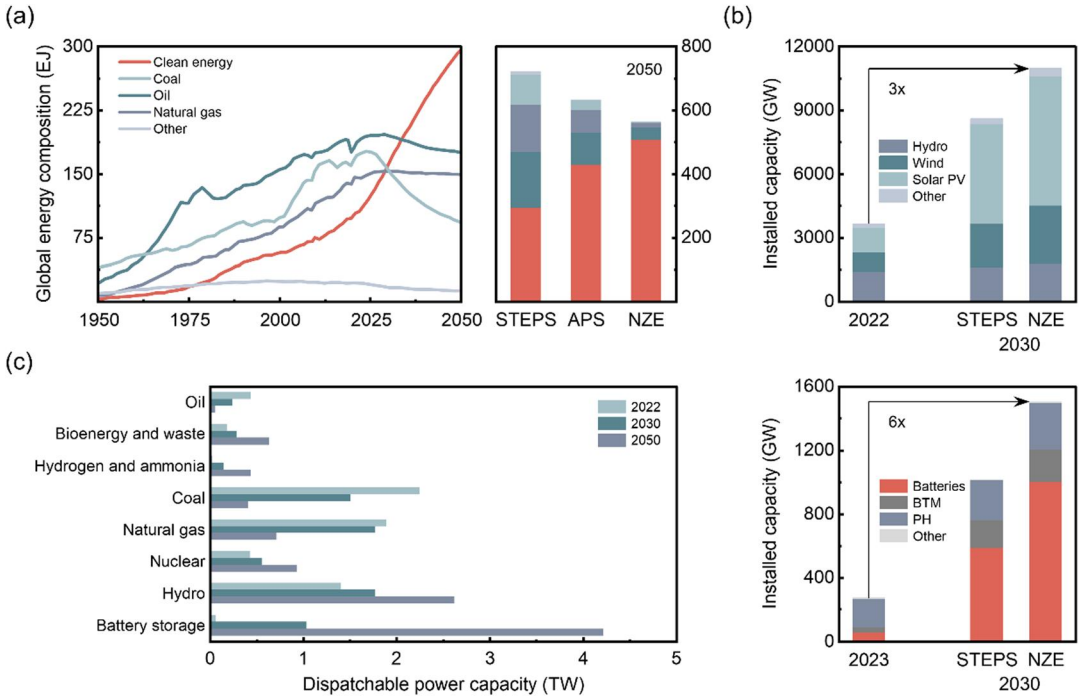

Fig. 1 The evolution of global renewable energy and dispatchable power capacity. (a) Global energy composition across scenarios (STEPS, stated policies scenario; APS, announced pledges scenario; NZE, net zero emissions by 2050 scenario) through 2050. EJ stands for exajoules. Oil, coal, and natural gas refer to unabated uses as well as non-energy use; clean energy includes renewables, modern bioenergy, nuclear, abated fossil fuels, low-emission hydrogen, and hydrogen-based fuels; residual categories cover routine biomass and non-renewable waste.[4] (b) Global installed renewable energy and energy storage capacity under STEPS and NZE scenarios (2022 baseline *vs.* 2030 projections). GW stands for gigawatts. Photovoltaics specified separately; supplementary renewables comprise bioenergy, geothermal, concentrating solar power, and marine; storage systems include compressed air, flywheel, and thermal technologies (hydrogen electrolyzers excluded). Batteries are the utility-scale batteries. BTM, behind-the-meter; PH, pumped hydro.[7] (c) Dispatchable power capacity by technology in the NZE scenario (2022 historical data, 2030/2050 projections). TW stands for terawatt. Hydrogen includes hydrogen and hydrogen-based fuel-fired power plants.[7] Data are obtained from ref. 4 and 7.

Agreement,[2] achieving net-zero CO_2 emissions by 2050 requires a major overhaul of energy infrastructure.[3] Under International Energy Agency's net zero emissions (NZE) scenario, clean energy is projected to supply 90% of global demand by mid-century (Fig. 1a).[4] However, the intermittency of renewables like wind and solar energy challenges grid stability, necessitating energy storage technologies.[5,6] Global storage capacity is expected to grow sixfold by 2030 compared to 2023, paralleling the tripling of renewable capacity (Fig. 1b).[7] Among storage technologies, batteries have rapidly improved, with battery costs reduced by 90% in less than 15 years.[8,9] Beyond capacity expansion, battery storage provides essential grid services through secure dispatchable capacity.[10] By 2050, batteries are projected to provide over 4 TW of installed capacity through continued innovation, becoming the dominant storage solution (Fig. 1c).[7]

Molecules lie at the heart of battery innovation, fundamentally shaping the performance, safety, and stability of modern electrochemical systems.[11–16] Among the various components, the electrolyte is considered as the "blood" of the battery, enabling ionic conduction while electronically insulating the electrodes.[17] Its molecular composition governs critical properties such as ionic conductivity,[18,19] electrochemical stability,[20,21] and flammability,[22,23] making electrolyte design central to battery performance. Most electrolytes, whether in liquid or gel form, are composed of small molecules or polymeric structures, whose physicochemical characteristics dictate the performance across diverse battery systems. For instance, lithium (Li)–sulfur (S) batteries,[24–26] despite their high theoretical energy density (2600 Wh kg^{-1}),[27,28] benefit from weakly solvating, encapsulating-polysulfide co-solvents,[29–32] such as hexyl methyl ether, to reduce side reactions between soluble polysulfides and

Qiang Zhang

Qiang Zhang is a professor at Tsinghua University. His current research interests are advanced energy materials, including dendrite-free lithium metal anode, lithium–sulfur batteries, and electrocatalysis, especially the structure design and full demonstration of advanced energy materials in working devices. He is the Editor-in-Chief of EES Batteries and an Advisor Editor of Angew. Chem.

the Li metal anode, extending cycle life from <60 to >140 cycles.[33] Aqueous zinc (Zn) metal batteries, prized for their intrinsic safety and low cost, are hampered by dendrite growth and hydrogen evolution.[34–37] By molecular engineering, Zn||Zn cells were enabled to operate stably for 21 000 cycles (700 h) under an extreme current density of 60 mA cm^{-2}, representing a technical breakthrough for practical Zn metal full cells.[38] Solid-state polymer batteries, recognized as highly promising next-generation energy storage devices due to their high energy density and enhanced safety profiles, nevertheless require meticulous monomer design and polymer matrix engineering to achieve practical ionic conductivity ($>10^{-3}$ S cm^{-1}) and stable interfacial compatibility.[39–42] Redox flow batteries offer cost-effective grid-scale storage with inherent scalability.[43,44] Molecular-engineered redox-active organic molecules achieve electrochemical stability >2 V (ideally over 4 V), facilitating the establishment of high-energy storage systems.[45]

In addition to electrolytes, electrode materials also depend on precise molecular and structural design to deliver optimal electrochemical performance.[46–49] Of particular interest are organic electrode materials, which afford environmental sustainability, synthetic tunability, and resource abundance as alternatives to conventional inorganic compounds.[50,51] Nonetheless, challenges such as low electrical conductivity and limited operating voltage are often encountered.[52,53] To address these limitations, structural motifs such as thiophene and furan rings are frequently used to enhance hole mobility in hole-transporting materials, while pyridine rings serve as key electron-withdrawing groups in electron-transporting analogues.[54–56] Furthermore, the operating voltage of organic electrodes can be finely tuned by modulating the energy of the lowest unoccupied molecular orbital (LUMO), highlighting the potential of molecular engineering to systematically optimize electrode function.[57] Rational molecular engineering holds significant promise for overcoming these limitations and achieving breakthroughs in electrode performance. Furthermore, other essential components such as binders[58–62] and separators[63,64] also benefit substantially from rational molecular design. Tailoring their molecular structures can improve mechanical strength, interfacial adhesion, thermal stability, and ion transport properties, thereby contributing to the overall performance and durability of the battery. Therefore, a unifying and versatile strategy of rational molecular design is essential to guide the exploration of electrolytes, electrodes, and auxiliary materials across a broad spectrum of battery chemistries.

1.2. AI for battery molecular discovery

While molecular design plays a foundational role in advancing battery technologies, the routine discovery approaches rooted in experimental screening remain slow and costly (Fig. 2).[65] The vastness of chemical space, combined with the need for precise control over molecular structure–property relationships, renders purely empirical methods inefficient for rapid innovation of advanced energy materials. In response, computational chemistry methods, such as density functional theory (DFT) calculations and molecular dynamics (MD) simulations, have provided powerful alternatives for probing molecular behavior and guiding rational design.[66–71] However, as the breadth of chemical space and the volume of data expand, computational methods become increasingly resource-intensive, requiring more time and incurring higher costs.[72,73]

The emergence of artificial intelligence (AI) offers a promising opportunity to transform the scientific paradigm.[74,75] Unlike routine approaches that rely on explicit physical modeling, AI systems learn from data to capture hidden correlations. The potential of this data-driven methodology has been widely recognized across the scientific community, as reflected in the 2024 Nobel Prizes, with the physics prize honoring foundational work in artificial neural networks (ANNs) by John Hopfield and Geoffrey Hinton, and the chemistry prize recognizing computational protein design and prediction breakthroughs by David Baker, Demis Hassabis, and John Jumper.[76,77] These milestones underscore the growing capability of AI in solving high-dimensional, nonlinear problems in chemistry and materials science.[78–80] With the continued expansion of computational power, AI is poised to deliver transformative breakthroughs in battery molecular discovery by significantly accelerating the discovery process (Fig. 3). Through accurate, computationally tractable surrogate models, AI bridges the gap between theory and experiment, thereby enabling the identification of novel, high-performance battery molecules.[81–84] Recent advances illustrate the potential in this field. For instance, Qiang Zhang and Xiang Chen's group at Tsinghua University has applied AI to accelerate

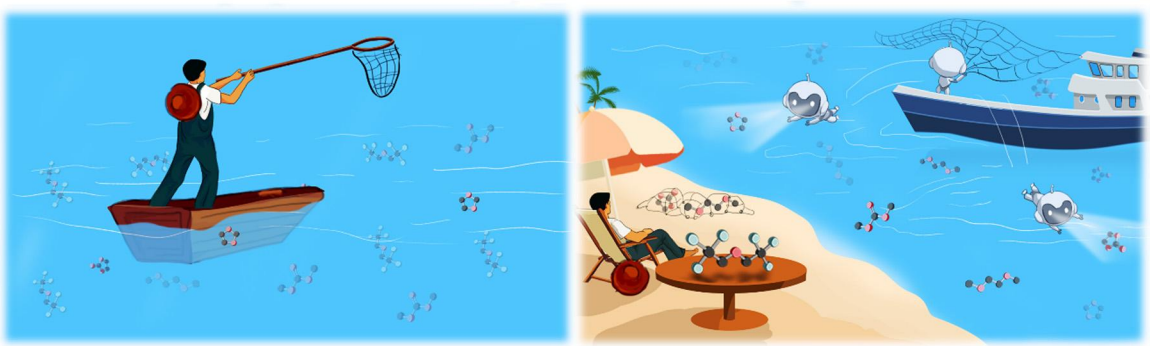

Fig. 2 Comparison of routine trial-and-error and emerging AI-assisted discovery of battery molecules.

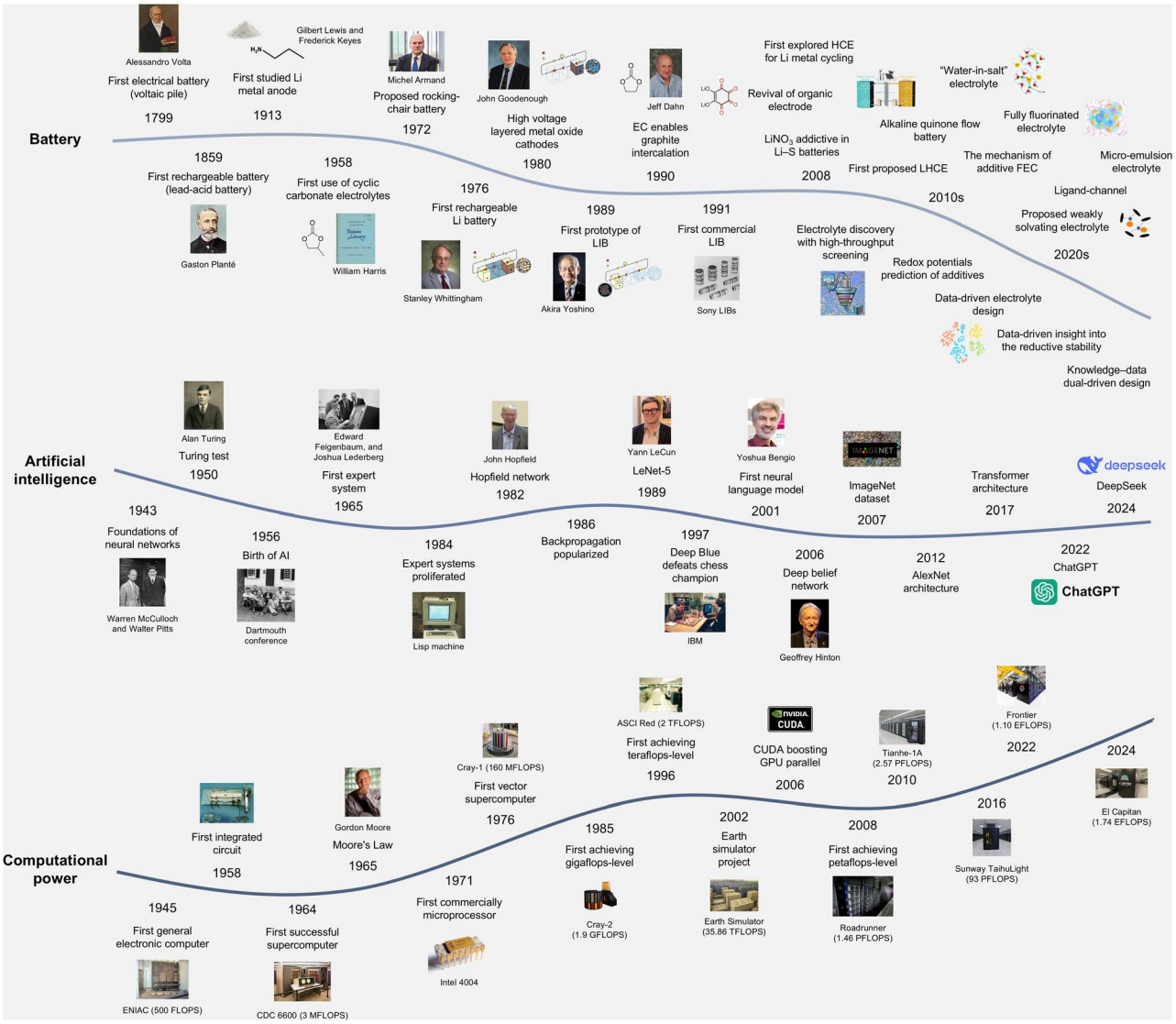

Fig. 3 A summary of the history of battery, AI, and computational power development. LIB, Li-ion battery; HCE, high-concentration electrolyte; LHCE, localized high-concentration electrolyte; EC, ethylene carbonate; FEC, fluoroethylene carbonate; GPU, graphics processing unit. Revival of organic electrode: reproduced with permission from ref. 89. Copyright 2008 Wiley-VCH. Electrolyte discovery with high-throughput screening: reproduced with permission from ref. 90. Copyright 2014 American Chemical Society. Alkaline quinone flow battery: reproduced with permission from ref. 91. Copyright 2015 American Association for the Advancement of Science. "Water-in-salt" electrolyte: reproduced with permission from ref. 92. Copyright 2015 American Association for the Advancement of Science. Fully fluorinated electrolyte: reproduced with permission from ref. 22. Copyright 2019 Springer Nature. Proposed weakly solvating electrolyte: reproduced with permission from ref. 93. Copyright 2020 Wiley-VCH. Data-driven insight into the reductive stability: reproduced with permission from ref. 85. Copyright 2023 American Chemical Society.

electrolyte design by identifying molecular features that influence reductive stability and constructing the data–knowledge dual-driven framework for electrolyte molecular design.[85–87] Yi Cui's team at Stanford University has leveraged AI techniques to establish structure–property relationships in electrolyte systems, enabling the design of molecular candidates that exhibit high Coulombic efficiency (CE).[88] These representative efforts demonstrate how AI can augment and accelerate molecular design pipelines, uncover mechanistic insights, and guide experimental exploration with higher precision.

While early studies have demonstrated the potential of AI in battery innovation, there remains a lack of a comprehensive review focusing on AI-driven molecule-level design strategies across the diverse landscape of battery chemistries. In response, this review affords a critical overview of how cutting-edge AI techniques are accelerating molecular innovation in batteries, bridging the knowledge gap and outlining opportunities at this emerging interdisciplinary frontier.

1.3. Scope of this review

In this contribution, the methodological and application innovations of the AI technique to battery molecules are comprehensively summarized and prospected, including molecular representation, AI models, molecular property prediction, and

molecular design for rechargeable batteries. Section 2 summarizes molecular representation strategies, encompassing foundational principles and computational methodologies with emphasis on multi-dimensional encoding schemes. These representation paradigms serve as the critical initial phase for AI-driven molecular discovery by transforming chemical entities into structured machine-readable inputs. Section 3 systematically categorizes AI modeling architectures, detailing key processes spanning data management, feature selection, model construction, and performance evaluation. The section further analyzes prevalent algorithms in molecular discovery including classical machine learning (ML, supervised/unsupervised learning methods), deep learning (DL, four basic architectures), and large language models (LLMs). Section 4 evaluates AI-powered property prediction for battery components, including redox potential, dielectric constant, ionic transport, and other fundamental physicochemical properties. Such computational capabilities provide robust tools for establishing structure–property relationships in battery molecules. Section 5 examines molecular design advancements through four approaches: interpretable ML (IML) for knowledge discovery, high-throughput virtual screening (HTVS), oriented molecular generation, and high-throughput experimentation (HTE). Case studies demonstrate paradigm-shifting applications of these methodologies in battery innovation. Finally, a summary concludes with a comprehensive analysis of AI-driven molecular discovery advancements in rechargeable batteries, coupled with critical perspectives on unresolved challenges and emerging research frontiers in the interdisciplinary domain.

2. Molecular representation

Molecular representation bridges the gap between molecular structures and AI methods in molecular innovation by converting chemical entities into numerical formats *via* feature engineering or representation learning. Effective representations require basic principles called the "2AI principles", including accurate, appropriate, invariant, and interpretable. This section systematically explores molecular representation paradigms, beginning with classical techniques such as one-hot encoding (OHE) and molecular fingerprints. Subsequent sections delve into advanced approaches, including graph-based representations that encode atomic connectivity and geometric descriptors preserving three-dimensional (3D) conformational information. The integration of expert-defined features with data-driven embeddings is emphasized, particularly in addressing challenges related to battery innovation. Finally, emerging trends such as hybrid architectures combining string or rule-based fragmentation with LLMs are discussed, highlighting their potential to advance high-throughput molecular discovery through chemically interpretable and computationally efficient representations.

2.1. Concept of molecular representation

In the field of organic chemistry, various molecular representations, such as bond-line structures, ball-and-stick models, and projection formulas, are frequently used to represent molecular structures.[94–96] These representations facilitate our understanding and manipulation of molecules in both experimental and theoretical studies. By providing an intuitive depiction of molecular structures, these representations enable accurate predictions and analyses of molecular interactions, reaction mechanisms, and other chemical properties. However, these traditional representations are not directly suitable for computer input. As a result, molecular representation methods have been explored to translate chemical molecular information into forms that can be understood and processed by AI models.[97–99]

Molecular representation refers to the abstraction of chemical molecular characteristics and their subsequent conversion into numerical data through specific encoding methods (Fig. 4).[100–103] The process serves as a crucial bridge between chemistry and computer science, forming the foundation for molecular AI research. In the development of battery-related molecules, the relationship between a fundamental molecular structure and its macroscopic properties, such as its impact on battery performance, is often complex. Molecular representation simplifies the complexity into a numerical format, enabling efficient analysis and prediction through AI models.

The process of converting molecular structures into numerical features is commonly referred to as feature extraction, which is a critical component of feature engineering and represents the first stage of an AI workflow.[104–107] The core objective of feature extraction is to draw on domain expertise to distill essential information that reflects the characteristics of a molecule from complex raw data.[83] Feature engineering has long played a pivotal role in traditional ML, where the quality of input data and the chosen features significantly influence the performance of the resulting models.[108]

With the advancement of DL technologies, feature engineering has progressively merged with the training process of AI models, giving rise to the research field of representation learning.[109,110] Representation learning leverages neural networks to automatically extract features from input data, thereby reducing the reliance

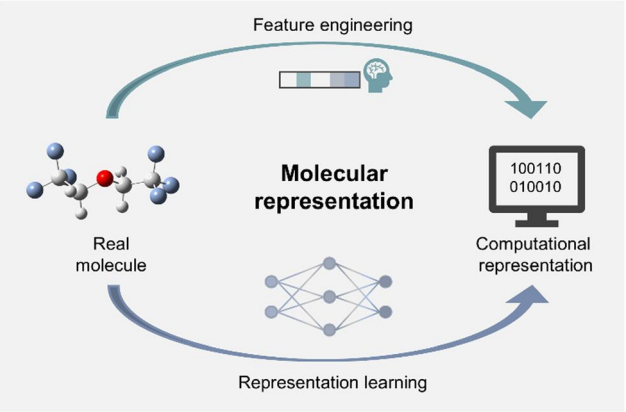

Fig. 4 The definition of molecular representation. Molecular representation refers to the numerical encoding of physical molecular entities for computational interpretation, which contains molecular feature engineering and representation learning.

on manually designed and selected features.[111–113] The approach enables AI to learn and extract meaningful features from vast datasets more efficiently. Within the representation learning framework, more fundamental molecular structural features can be directly utilized, allowing models to deeply explore structure–property relationships, which increasingly positions representation learning as a pivotal direction for accelerating molecular design with AI. However, while representation learning can reduce the reliance on pre-defined molecular features, the design of molecular representations should be aligned with the architecture and optimization of neural network models to ensure that molecular characteristics are effectively captured.[114,115]

2.2. Basic principles of molecular representation

An effective molecular representation method should meet the four fundamental requirements called the "2AI principles" to ensure high efficiency and practicality at the intersection of chemistry and AI (Fig. 5).

2.2.1. Accurate. An accurate molecular representation is defined by its ability to capture the breadth of chemical space while simultaneously distinguishing subtle variations among different molecules.[101]

First, representations are expected to encode complete molecular structures.[116] Incomplete encodings can deprive AI models of the full data context required for precise task execution. At the same time, extraneous redundancy within the representation should be minimized, since an overabundance of descriptors can impose an unnecessary learning burden and degrade model performance.[117] Second, subtle distinctions between closely related molecules must be faithfully preserved. For example, tautomeric shifts often induce dramatic changes in molecular properties and continue to challenge current representation schemes.[118] If the chosen descriptors fail to reflect these nuanced differences, the ability of models to learn the unique characteristics of each molecule will be compromised, impairing their capacity to perform complex chemical tasks.

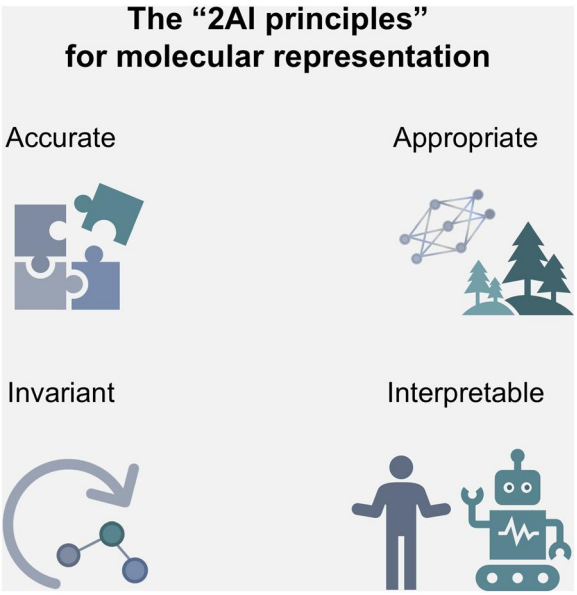

Fig. 5 The "2AI principles" of molecular representation.

2.2.2. Appropriate. An appropriate molecular representation is one, in which descriptor selection is tailored to the specific application context, as no single scheme has yet proven to be universally optimal.[97,119] Considerations such as dataset size, the choice of AI algorithm, knowledge-driven or generalization-driven objectives, and which descriptors will yield the best performance.

When learning tasks involve relatively small sample sizes (on the order of a few hundred), the adoption of low-dimensional molecular representations is generally considered appropriate.[120] For example, Li *et al.*[121] selected 37 descriptors from an initial pool of 199, and overall classification accuracy across six models was raised from 46.8%–79.1% to 71.0%–83.7%. In addition, Okamoto *et al.*[122] predicted the redox potential of Li-ion battery (LIB) additives using fewer features selected through importance analysis, achieving comparable performance to using all features while enhancing model interpretability. On the other hand, representation learning is generally more suitable when working with large datasets. Fang *et al.*[112] trained the molecular representation learning model GeoGNN on 20 million data points, achieving state-of-the-art on 14 of 15 molecular property prediction benchmarks. The size of the dataset is crucial for representation learning models to stand out.[123]

The choice of representation is often aligned with the downstream AI architecture. Traditional ML algorithms typically employ molecular fingerprints and descriptors, while DL architectures favor graph-based representations for graph neural networks (GNNs) and string-based encodings for sequence models. In addition, choosing an appropriate molecular representation method also requires considerations from different application scenarios. Domain knowledge-intensive tasks favor expert descriptors and fingerprints encoding established chemical principles, while generalization-driven applications benefit from geometric DL that integrates molecular graphs or coordinates through pretrain and fine-tune frameworks.

2.2.3. Invariant. The process of molecular representation learning typically relies on AI models to automatically extract features. During modeling, it is crucial to carefully consider the invariance and equivariance of the model, as these govern scalar properties such as potential energy and vector quantities such as atomic forces.[124,125] Embedding these symmetries incorporates physical priors, thereby enhancing model training and improving accuracy.[126–128]

In the context of molecular modeling and neural network training, invariance refers to physical quantities that remain unchanged under certain transformations, while equivariance implies that certain physical quantities change in a predictable and consistent manner under specific transformations.[129–131] Four fundamental transformations dictate symmetry considerations:

(1) Translation: both potential energy and force magnitudes remain unchanged, indicating that the molecular system exhibits the same potential energy and forces at any position in space.

(2) Rotation: potential energy maintains rotational invariance while forces demonstrate rotational equivariance; in the latter case, vector directions rotate with molecular orientation while magnitudes persist.

(3) Permutation: permutation refers to changes in the order or position of atoms within a molecule. Forces exhibit equivariance in this context, while such permutations do not alter the potential energy, since the interatomic interactions remain unchanged.

(4) Reflection: chirality-related mirror transformations maintain potential energy but symmetrize force vectors. Due to the limited application of chiral molecules in battery systems, this transformation is generally not emphasized in most contributions.[132]

In practice, translational invariance is relatively straightforward to achieve, as the model does not rely on the absolute coordinates of the molecule but rather on the relative coordinates between atoms. Permutation invariance or equivariance can also be implemented naturally within neural networks, since the model is designed to be independent of the input order, therefore avoiding input-order-sensitive architectures such as RNNs along the atomic dimension. Consequently, achieving rotational equivariance is a crucial consideration in constructing molecular representation networks.

2.2.4. Interpretable. The interpretability of molecular representations is expected to significantly enhance the understanding of structure–property relationships in molecular innovation. Chemically meaningful representations are supposed to adhere to established chemical principles and afford actionable insights into the decision-making processes of predictive models. Such interpretability helps ensure that AI predictions remain consistent with domain knowledge while maintaining the level of transparency necessary for experimental validation and mechanistic hypothesis development. In this way, interpretable representations facilitate the connection between AI outputs and scientific reasoning, allowing it to be assessed whether model predictions arise from chemically valid patterns or from artifacts within the data.

Different representation methods possess varying degrees of intrinsic interpretability. The routine low-dimensional representations, such as expert descriptors including dipole moment and binding energy, are characterized by explicit encoding of chemically intuitive features, which confers inherent interpretability. Molecular fingerprinting techniques, exemplified by the molecular access system (MACCS) keys, demonstrate interpretability through their rule-based construction and chemically intuitive feature encoding. These approaches utilize rule-based systems to generate human-readable numerical embeddings that align with fundamental chemical concepts. Although transparent structure–property relationships can be established through these representations, their reliance on predefined feature spaces tends to limit predictive accuracy when applied to complex material behaviors.

In contrast, DL approaches employ automated feature extraction from molecular graphs and 3D conformations, often achieving higher predictive performance. However, interpretability is diminished due to the abstract nature of high-dimensional latent representations. To address this limitation, emerging methodologies incorporate physics-informed hybrid architectures, in which domain knowledge, such as symmetry constraints and quantum mechanical (QM) descriptors, is systematically integrated into neural network models. These advanced representations are designed to combine the expressive capacity of DL with chemical consistency, thereby promoting a balance between predictive accuracy and scientifically meaningful interpretability.

2.3. Categories of molecular representation methods

Molecular representation design remains task-dependent, requiring careful alignment between chemical knowledge, computational constraints, and scientific objectives. Representation methods develop across one-dimensional (1D), two-dimensional (2D), 3D, and four-dimensional (4D) scales:

(1) 1D sequential: OHE and strings prioritize computational efficiency but lack spatial awareness.

(2) 2D topological: molecular fingerprints and graph structures encode connectivity patterns.

(3) 3D geometric: QM fields and coordinates preserve stereochemical information.

(4) 4D dynamic: MD trajectories incorporate temporal evolution through atomic position time series.

Furthermore, from a methodological perspective, molecular representations can be categorized into two principal paradigms: fixed (expert-designed descriptors with explicit semantics) and differentiable (neural network-generated latent embeddings) representations. Modern hybrid architectures strategically combine these paradigms, as exemplified by GNNs that employ atom-type embeddings for initialization and equivariant Transformers that construct 3D conformations from simplified molecular input line entry system (SMILES) inputs.[133–138] The multi-scale integration framework advances beyond conventional feature engineering through systematic fusion of physical principles with data-driven learning mechanisms. In this section, the most commonly used molecular representation methods are introduced (Fig. 6).

2.3.1. One-hot encoding. OHE remains a foundational technique in molecular representation, systematically converting molecular features into binary vectors.[139,140] Each predefined chemical attribute, such as atom type, functional group, or substructure, occupies a unique vector index activated (represented by 1) upon presence or deactivated (represented by 0) otherwise. Implementation involves domain-informed vocabulary construction, fixed-position feature mapping, and sparse vector generation. For instance, a molecule containing benzene rings and double bonds can activate corresponding indices in a [hydroxyl, amino, benzene ring, double bond] vocabulary, yielding [0,0,1,1].

Despite its simplicity, OHE exhibits critical limitations.[141–144] Firstly, vocabulary size linearly scales vector dimensions, causing computational inefficiency and dimensional curses at thousand-feature scales. Secondly, binary activation discards feature frequency and topological relationships, compromising chemical fidelity. Thirdly, sparse representations exacerbate overfitting risks through redundant zero inflation. Therefore, advanced frameworks address these constraints by integrating OHE as a foundational component within DL architectures.

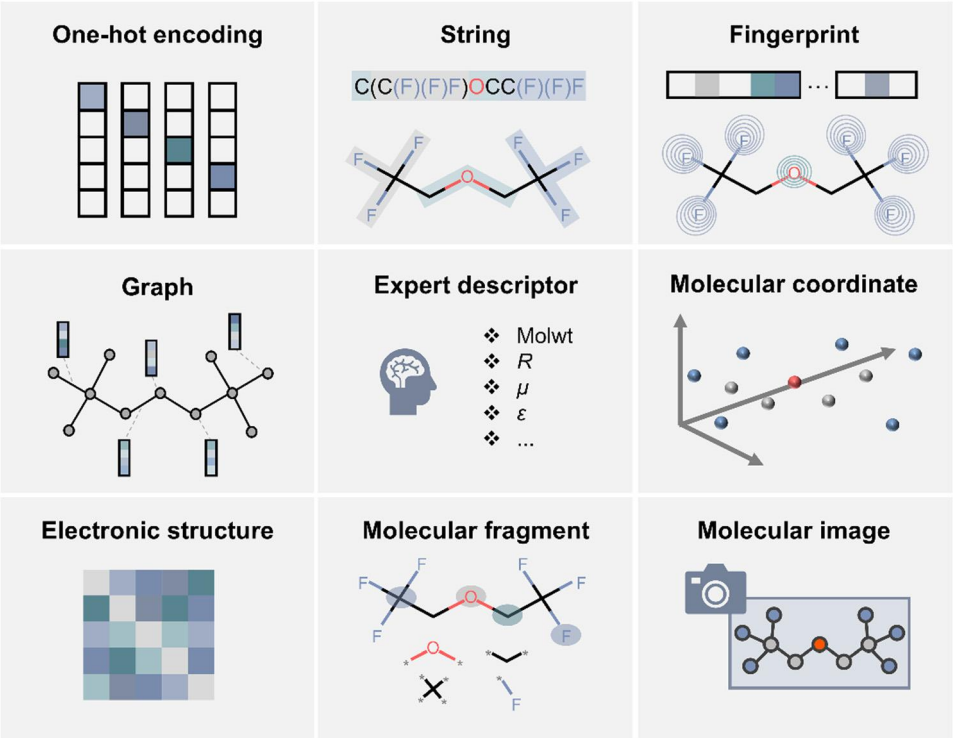

Fig. 6 The categories of molecular representation methods.

Gilmer *et al.*[145] demonstrated this synergistic integration through message passing neural networks (MPNNs), where OHE initializes atom-type embeddings for fundamental elements and quantizes bond distances into discrete, distance-binned one-hot vectors. Large-scale systems like AlphaFold3[135] further demonstrate the versatility of OHE in encoding biomolecular entities, sequence alignments, and spatial relationships.

2.3.2. String. The development of molecular representation systems began with the Wiswesser line notation in 1949,[146–149] which pioneered linear chemical encoding through alphanumeric combinations but faced obsolescence due to syntactic complexity. Subsequently, the SMILES, introduced by David Weininger in 1988,[150] revolutionized the field by prioritizing human readability and computational compatibility. SMILES employs element symbols for atoms, implicit hydrogen deduction, and explicit bond notation (single bonds omitted, double/triple bonds marked as =/#). Cyclic structures are encoded numerically, while branching uses nested parentheses. Aromatic systems alternate between explicit bond annotation (*e.g.*, C1═CC═CC═C1) and lowercase atomic symbols (*e.g.*, c1ccccc1), balancing structural clarity with compactness.

SMILES exhibit inherent non-uniqueness due to graph traversal variations during generation, a property later exploited for data augmentation *via* randomized atom ordering.[151–155] To address database standardization needs, canonical SMILES emerged, enforcing uniqueness through deterministic rules like Morgan identifier sorting.[156–160] Subsequent advancements tailored SMILES for AI integration: for example, self-referencing embedded strings[161,162] introduced grammatical constraints to ensure valid structure generation, DeepSMILES[163,164] optimized syntax for neural network efficiency, and SMILES arbitrary target specification (SMARTS)[165–167] extended functionality as a substructure query language. SMARTS enhances SMILES with logical operators, environment descriptors, and bond qualifiers, enabling precise pattern matching.

Complementary identifiers address specialized requirements. The International Union of Pure and Applied Chemistry[168–170] nomenclature provides systematic structural descriptions but lacks algorithmic friendliness. CAS registry numbers[171,172] enable compound tracking without structural insights. International chemical identifier (InChI)[173–175] and its hashed derivative InChIKey[176,177] employ layered encoding (molecular skeleton, charge, stereochemistry) for cross-database compatibility, trading readability for standardization. Database-specific identifiers (PubChem CID,[178] ChemSpider ID,[179] Material ID[180,181]) prioritize rapid indexing at the cost of cross-platform interoperability. The ecosystem of string representations balances human interpretability, computational efficiency, and application-specific needs across chemical research and informatics.

2.3.3. Molecular fingerprint. Molecular fingerprint is a molecular representation that encodes specific structural and/or physicochemical features of a molecule into a numerical vector.[182,183] These encodings, typically as binary presence indicators or integer frequency counts, enable machine-readable chemical data essential for similarity searching,[184] virtual screening,[185] and AI model development.[186,187] Early rule-based systems like MACCS[188] employed expert-curated substructure dictionaries, translating predefined features into binary bit strings.

Initially, MACCS contained 960 expert-designed structural keys, followed by the release of a publicly available 166-bit reduced version (public MACCS keys), which can be efficiently computed using open-source tools like RDKit[189] and OpenBabel.[190] Subsequent expansions, exemplified by the 881-bit fingerprint of PubChem,[191] systematically cataloged pharmacophores and topological motifs. Such preset fingerprints have been widely used for molecular similarity searching and substructure matching with the Tanimoto coefficient.[192] The Tanimoto coefficient quantifies molecular similarity by comparing the proportion of common activated bit positions (*i.e.*, the ratio of intersection to union of bits) between two binary fingerprints.[193] However, such fingerprints inherently limit feature space to manual definitions, potentially missing novel structural relationships.

Preset fingerprints faced inherent constraints in representing novel chemical systems, driving innovation in topology-driven encoding. The extended connectivity fingerprint (ECFP)[194,195] was introduced as a significant implementation based on the atomic environment concepts proposed by Morgan.[156] ECFP dynamically encodes molecular topology through three stages: atomic initialization (assigning identifiers based on atomic type/valence/adjacent environment), iterative neighborhood expansion, and hashing with redundancy removal. The process generates adaptive features without a predefined dictionary, capturing atomic-to-global structural patterns. According to the different initial atomic identifiers, the fingerprints generated by ECFP can be divided into standard ECFP fingerprints (hash atomic physical properties) and functional-class fingerprints (map to pharmacophore functional codes). Additionally, depending on the range of the expansion radius, the most applied ECFP*x* series includes ECFP4 and ECFP6. By adjusting the expansion radius, a balance between feature resolution and computational efficiency can be further optimized.

Aside from the dictionary-based fingerprints (including MACCS,[188] PubChem fingerprint,[196] and SMILES fingerprint[197]) and circular fingerprints (including ECFP,[195] FCFP,[198] MinHash fingerprint,[199] Molprint2D,[200] and Molprint3D[201]) discussed above, topological fingerprints (including atom pair fingerprint,[202] atom pair fingerprint extended with atom properties fingerprint,[203] and topological torsion fingerprint[204]) and shape-based fingerprints (including rapid overlay of chemical structures[205] and ultrafast shape recognition[206]) are also widely employed in molecular representation.

2.3.4. Graph. One molecule naturally forms a graph with atoms as nodes (vertices) and bonds as edges, typically modeled as an undirected graph structure in mathematics.[207,208] In formal terms, a molecule can be represented as $\mathbf{G} = \langle \mathbf{V}, \mathbf{E} \rangle$, where the vertex set $\mathbf{V} = \{v_1, v_2, \ldots, v_n\}$ commonly includes all heavy-atom (non-hydrogen) atoms and the edge set $\mathbf{E} = \{e_1, e_2, \ldots, e_m\}$ describes the connectivity between atoms. The vertex attribute matrix encodes physical and chemical features such as atom type, hybridization state, and charge, and the edge attribute matrix records bonding information such as bond order, bond length, and aromaticity.[209–212]

Early developments in chemical graph theory focused on the molecular topological features. The 1940s saw the introduction of the Wiener index,[213,214] which describes molecular structures through atomic path sums, followed by the Hosoya index,[215,216] which has been reported to correlate alkane boiling points with bond matching patterns, and the Randić index,[217–219] which is suitable for measuring the extent of branching in the carbon-atom skeletons of saturated hydrocarbons. These indices underpinned quantitative structure–property relationship (QSPR) models, establishing foundational links between molecular topology and macroscopic behavior.[220–223]

In computer implementations, the storage and manipulation of molecular graphs rely on classical graph data structures.[224,225] The adjacency matrix uses a 2D array $\mathbf{A}_{nn}$ to precisely describe the atomic connectivity, where $\mathbf{A}_{ij}$ represents the bond order between atom i and atom j. On the other hand, the adjacency list records each atomic neighboring node and bond attributes in a linked list format, which enhances the storage efficiency for long-chain molecules or sparse structures such as branched polymers. Once a molecular graph is represented in a computer, graph algorithms such as depth-first search,[226–228] breadth-first search,[229–231] shortest path algorithms,[232–234] and subgraph isomorphism algorithms[235–237] can be applied to address topological problems on the molecular graph. With DL technologies emerging, the representation of molecular graphs has naturally transitioned to the computational paradigm of GNNs.[238–240] GNNs implement a message passing framework (*e.g.*, MPNN) to simulate the local interactions between atoms, enabling the automatic extraction of underlying chemical patterns.[241] The technical details and cutting-edge applications of GNNs will be systematically discussed in subsequent sections.

2.3.5. Expert descriptors. The construction of expert descriptors essentially involves translating domain knowledge into quantifiable molecular features, with the design closely aligned with the physical and chemical mechanisms underlying the target property.[242–245] Property-driven feature engineering is particularly crucial in exploring battery molecules.[246–249] The construction process is usually guided by the intuition of domain experts and typically relies on DFT calculations, MD simulations, or the in-depth analysis of experimental characterization data. For instance, Allam *et al.*[250] established an ANN utilizing descriptors including electron affinity, the highest occupied molecular orbital (HOMO), LUMO, HOMO–LUMO gap, and atom counts, to predict the redox potentials of organic electrode molecules.

Since expert descriptors often encompass parameters with different units and scales, it becomes essential to preprocess the data and scale the data to a common range to ensure the robustness of the model. Z-score normalization transforms each feature into a distribution with a mean of 0 and a standard deviation of 1, making it suitable for Gaussian-distributed data.[251] The formula for Z-score normalization is

$$x' = \frac{x - \mu}{\sigma} \tag{1}$$

where x' is the normalized value, x the original value, μ the mean of the feature, and σ the standard deviation of the feature. On the other hand, min–max normalization linearly

maps the features to the [0, 1] range, which is especially useful for parameters with well-defined boundaries.[252] The formula for min–max normalization is

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{2}$$

where $x_{\min}$ is the minimum value of the feature and $x_{\max}$ the maximum value of the feature.

2.3.6. Molecular coordinate. Molecular 3D coordinate representation describes the geometric configuration and stereochemical features of a molecule through the atomic positions in Cartesian space, along with their connectivity.[253,254] Molecular coordinate not only allows for the static depiction of bond lengths, bond angles, and dihedral angles, but also enables the dynamic capture of conformational changes through multiple coordinate frames.[255,256]

The commonly used 3D coordinate representation formats are designed with distinct approaches, considering information density and specific application scenarios. The XYZ format,[257] as the simplest representation method, records only atomic types and their spatial coordinates, making it suitable for the rapid storage and exchange of small-molecule conformations. However, the XYZ format lacks bond connectivity information and requires external algorithms to reconstruct the molecular topology. The Structure Data File format[258] further integrates multiple molecular data blocks and custom physicochemical property fields, supporting the efficient storage and retrieval of large molecular libraries. The Protein Data Bank format[259,260] includes atomic coordinates, sequence information, crystallographic parameters, and experimental metadata, and is a core data carrier in structural biology research.

2.3.7. Electronic structure. Electronic structure representation provides a mathematical formalization of the electronic distribution characteristics within a molecular system, offering input features for ML models that combine physical interpretability with computational robustness. The Coulomb matrix[261] is a classical descriptor for electron interactions. For a molecule containing N atoms, the Coulomb matrix $\mathbf{C} \in \mathbb{R}^{N\times N}$ has matrix elements $\mathbf{C}_{ij}$ given by

$$\mathbf{C}_{ij} = \begin{cases} 0.5Z_i^{2.4}, & \text{if } i = j, \\ \dfrac{Z_i Z_j}{\|\mathbf{r}_i - \mathbf{r}_j\|}, & \text{if } i \neq j. \end{cases} \tag{3}$$

where Z_i and Z_j represent the atomic numbers (nuclear charge) of atoms i and j, respectively, $\mathbf{r}_i$ and $\mathbf{r}_j$ the corresponding 3D atomic coordinate vectors, and $||\cdot||$ the Euclidean norm. The diagonal elements are empirically corrected by the 2.4$^{\text{th}}$ power of the atomic number Z. The off-diagonal elements describe the Coulomb repulsive potential between pairs of atoms, therefore encoding the electrostatic interaction network within the molecule as a symmetric matrix. Despite the rotational and translational invariance of the Coulomb matrix, its numerical values are highly sensitive to the atom indexing order.[262] Many methods have been proposed to solve the problem, such as random and sorted Coulomb matrices,[263] employing the Wasserstein norm, rather than Euclidean or Manhattan norms,[264] and many-body tensor representations.[265] Furthermore, the concept of the Coulomb matrix is further extended to describe broader interactions, such as the Ewald sum matrix,[266] Sine matrix,[266] and Bag of Bonds model.[267]

2.3.8. Molecular fragment. Molecular fragment decomposes molecules into chemically meaningful substructures, such as functional groups, scaffolds, and reactive motifs, to enable modular analysis and design.[268–271] These fragments are typically generated through structure-based methods, in which invariant core units such as scaffolds are identified by analyzing atomic connectivity and bonding patterns.[272,273] For instance, Degen *et al.*[274] proposed the breaking of retrosynthetically interesting chemical substructures (BRICS), a rule-based fragmentation method employing 16 retrosynthetically inspired cleavage rules to partition molecules into chemically relevant motifs. Diao *et al.*[275] extended BRICS, introducing user-definable parameters and graph-based decomposition to generate diverse fragments. In addition, fragments can also be generated using data-driven approaches, in which fragment vocabularies are learned from chemical databases.[276–278] For instance, Li *et al.*[138] developed SMILES pair encoding, a data-driven tokenization algorithm that augments atom-level encodings with SMILES substrings learned from extensive chemical datasets.

Fragment-based methods remain limited in capturing global molecular topology and rely on predefined fragment libraries, constraining their adaptability to novel chemical spaces.[279–281] Synergizing with LLMs offers transformative potential, as treating fragments as lexical units (tokens) enables LLMs to model substructural relationships and generate synthetically feasible molecules beyond existing libraries.[282–285]

2.3.9. Molecular image. Molecular image representation encodes chemical structures into visual formats.[286–292] Images range from 2D line-angle drawings to multi-view 3D projections and property-enhanced visualizations like electron density isosurfaces or electrostatic potential heatmaps.[293–295] Such representations enable convolutional neural networks (CNNs) to implicitly learn spatial topology and stereochemical constraints through end-to-end training.[296–299] However, molecular image representations suffer from structural information loss during rasterization (*e.g.*, omitted stereochemical details or electronic properties), resolution and viewpoint sensitivity, and non-standardized rendering across studies, motivating careful dataset curation and normalization protocols.[300–303]

2.4. Molecular toolkits

A diverse ecosystem of software libraries has emerged to support both the generation of chemical structures and the derivation of molecular representations for modeling. Table 1 summarizes the commonly used chemical toolkits.

As discussed in previous parts, it is natural to use a graph to represent molecular structures. Graph-theoretic libraries enable rapid prototyping of custom molecular graph algorithms and workflows. NetworkX provides a flexible framework for creating, manipulating, and analyzing arbitrary graphs in Python, making it suitable for encoding bond connectivity and for performing

Table 1 Molecular toolkits

Toolkit	Description	Ref.
AlvaDesc	A commercial application for calculating almost 6000 molecular descriptors, molecular fingerprints, and user-defined structural patterns	304
BlueDesc	A ready-to-use, General Public License-licensed Java tool computing 174 3D molecular descriptors *via* integrated JOELib2 and CDK libraries, converting an MDL SD file into ARFF or LIBSVM format for QSAR/QSPR modeling *via* command-line execution	305
CDK	An open-source modular Java library for cheminformatics and bioinformatics, supporting molecular parsing, manipulation, descriptor calculation, fingerprint generation, and substructure search	306
ChemDes	A free web-based platform for calculating 3679 molecular descriptors across 59 fingerprint types, with auxiliary tools for format conversion, MOPAC optimization, and fingerprint similarity analysis	307
chem^{f}	An open-source, purely functional cheminformatics toolkit written in Scala, providing immutable molecular representations, SMILES parsing, and graph-based operations	308
Chemkit	An open-source C++ toolkit library for cheminformatics, molecular modeling, and molecular visualization	309
ChemmineR	An open-source R package for analyzing drug-like small-molecule datasets, offering physicochemical property prediction, structural similarity searching, clustering, and visualization	310
ChemoPy	An open-source Python package, including 1135 features and seven molecular fingerprint systems, with optional semi-empirical MOPAC integration for extensive 3D descriptor computation	311
ChemSAR	An online access pipelining platform, offering validation and standardization of structures, computation of 783 descriptors and ten fingerprint types, and stepwise model generation	312
ChemTools	An open-source Python library for interpreting quantum chemistry outputs, translating electronic structure theory results into chemical descriptors	313
Cinfony	An open-source Python package for creating, manipulating, and studying complex networks with efficient graph algorithms can be used to generate molecular structures as graph representations	314
Daylight	A commercial cheminformatics toolkit by Daylight, providing object-oriented APIs for retrieving chemical information and integrated wrappers for C/C++ with Java	315
DeepChem	An open-source Python library democratizing DL, providing graph-based featurizers, dataset utilities, and neural network models	316
Dragon	A commercial application developed by Talete SRL for calculating over 5000 molecular descriptors, including atom types, functional groups, fragment counts, topological, and geometrical descriptors	317
E-Dragon	An electronic remote version of the commercial DRAGON descriptor engine, enabling remote calculation of over 1600 molecular descriptors from SMILES, SDF, or MOL2 files	318
igraph	An open-source C library with Python, R, and Mathematica interfaces for high-performance network analysis that can be used to generate molecular structures as graph representations.	319
Indigo	An open-source cheminformatics library in C++, providing SMILES canonicalization, reaction processing, and R-group deconvolution, with cross-language bindings (Python, Java,.NET, R, WebAssembly)	320
jCompoundMapper	An open-source Java library and command-line tool built on CDK for executing molecular decompositions, computing molecular fingerprints, and exporting fingerprints to ML formats	321
Mold2	An open-source command-line tool for fast calculation of 777 molecular descriptors from 2D chemical structures	322
MoleculeKit	An open-source Python library providing object-oriented classes and methods for reading, writing, visualizing, and manipulating biomolecular structures and trajectories	323
Mordred	An open-source Python toolkit providing rapid calculation of 1800+ molecular descriptors *via* command-line, web, or Python interfaces, supporting parallel processing and cross-platform deployment	324
NetworkX	An open-source Python package for creating, manipulating, and studying complex networks with efficient graph algorithms that can be used to generate molecular structures as graph representations	325
Open Babel	An open-source C++ library and toolkit with Python, Java,.NET, and Ruby bindings for interconverting over 110 chemical file formats, performing molecular modeling, descriptor calculation, and fingerprint generation	326
OpenChemLib	An open-source Java-based cheminformatics framework, providing molecular representation, substructure and reaction search, and GUI components for embedding chemical editors and viewers	327
OPSIN	A freely available software that interprets systematic IUPAC nomenclature for chemistry and biochemistry and converts it into chemical structures reported as SMILES, InChI, and CML	328
PaDEL	An open-source Java-based application leveraging CDK for multithreaded calculation of 797 molecular descriptors (663 1D/2D, 134 3D) and ten fingerprint types, featuring both graphical user interface (GUI) and command-line interfaces, and supporting over 90 file formats	329
Pybel	An open-source Python wrapper for the Open Babel API, providing Pythonic access to molecular objects, file I/O, descriptor and fingerprint calculation, SMARTS matching	330
PyDPI	An open-source Python molecular informatics platform integrating chemoinformatics and bioinformatics, computing structural descriptors and seven fingerprint systems	331
Rcpi	An open-source R package integrating bioinformatics and chemoinformatics, offering functions for computing sequence- and structure-based descriptors and fingerprints	332
RDKit	An open-source cheminformatics toolkit written in C++ and Python, offering comprehensive functionality for 2D/3D molecular operations, descriptor and fingerprint generation, and substructure searching	189
SMSD	A Java-based library for identifying maximum common subgraphs and substructures between small molecules, enabling precise chemical similarity and substructure searches	333
Surge	An open-source C-based chemical structure generator employing the Nauty package to enumerate all non-isomorphic constitutional isomers of specified molecular formulas	334

graph traversals or subgraph mining within chemical datasets.[325] Similarly, the igraph library offers high-performance graph data structures and algorithms, supporting both Python and R interfaces, and its C-core ensures scalability when enumerating large combinatorial chemical spaces.[319] Beyond general-purpose graph libraries, dedicated enumeration tools such as Surge have been developed for exhaustive isomer and scaffold generation.[334] Surge implements efficient, rule-based algorithms to enumerate structural isomers under valence and atom-typing constraints, facilitating the systematic exploration of chemical space in virtual screening campaigns.

Besides molecular structure generation, a variety of toolkits enable the computation of molecular fingerprints and descriptors. For instance, RDKit offers an extensive suite of fingerprint algorithms, including MACCS and ECFP, as well as physicochemical descriptors (*e.g.*, topological, electronic, and geometric properties), and cheminformatics utilities for tasks such as SMARTS-based substructure searching and reaction enumeration.[189] Open Babel provides command-line and library interfaces for interconverting molecular file formats and computing a range of fingerprints.[326] The chemistry development kit (CDK), written in Java, implements many descriptor calculators while exposing a modular application programming interface (API) for integration into custom pipelines.[306] In addition, recent advances in neural networks have been integrated into molecular modeling toolchains. DeepChem provides TensorFlow- and PyTorch-based implementations of graph convolutional networks (GCNs), MPNNs, and Transformers for chemistry.[316] Its pipelines enable end-to-end learning of continuous molecular embeddings directly from graph or SMILES inputs. Generative deep models, as supported in DeepChem, facilitate the inverse design of molecules by learning latent spaces that capture meaningful chemical variations.

By integrating these chemical toolkits into coherent workflows, researchers are now equipped with turnkey solutions that greatly simplify the application of AI in molecular science, thereby empowering chemists to leverage AI-driven methods.

3. AI models

AI algorithms play a pivotal role in accelerating molecular discovery by systematically identifying patterns from complex molecular structures, thereby significantly enhancing the exploration of new molecules. In this section, the foundational components of AI methodologies are introduced, encompassing aspects of data management, feature selection, model construction, performance evaluation, and practical applications. Subsequently, both prevalent ML methods and advanced DL frameworks are discussed. Special emphasis will also be placed on emerging LLMs, highlighting their innovative applications and illustrating representative case studies within molecular discovery.

3.1. Basic concepts of AI

The conceptual foundations of AI trace back to the mid-20th century, with pivotal milestones shaping its theoretical and practical evolution. The concept of AI can be traced back to the philosophical question "Can machines think?" posed by Alan Turing. In 1950, he laid the foundational groundwork for machine intelligence by proposing the Turing test as a criterion to assess whether machines could exhibit behavior indistinguishable from that of humans.[335] Then, the term of *artificial intelligence* was formally coined at the 1956 Dartmouth Conference, where John McCarthy, Marvin Minsky, and other pioneers outlined a research agenda to explore "how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves".[336] The conference marked the birth of AI as a distinct interdisciplinary field, integrating computer science, mathematics, and cognitive science.

At its core, AI encompasses systems designed to perform tasks requiring human-like intelligence, including reasoning, knowledge representation, problem-solving, perception, and learning. Early AI focused on symbolic approaches, relying on rule-based systems and logic. Over time, the field shifted toward data-driven methodologies. A transformative advancement emerged with ML, a subfield where algorithms autonomously identify patterns in data to make predictions or decisions without explicit programming. The learning process involves training models on labeled or unlabeled datasets, optimizing parameters to minimize errors between predictions and ground truth, a mathematical framework formalized by the concept of empirical risk minimization. Supervised learning, unsupervised learning, and reinforcement learning represent three primary paradigms, each addressing distinct challenges such as regression and classification, clustering, and sequential decision-making.[337]

The rise of DL in the 21st century revolutionized ML by leveraging hierarchical ANNs inspired by biological neural systems. Unlike shallow models, deep neural networks employ multiple layers to progressively extract high-level features from raw data, enabling breakthroughs in computer vision (CV),[338–340] natural language processing (NLP),[341–343] and robotics.[344–346] In CV, AI-driven systems now surpass human performance on certain benchmarks in tasks such as image classification and object detection,[347–349] while in NLP, LLMs enable unprecedented capabilities in semantic understanding and context-aware generation.[350–352] Similarly, AI-powered robotics integrates perception, decision-making, and control to achieve autonomous operation in dynamic environments.[353–355] The development of backpropagation algorithms[356] and computational advancements (*e.g.*, graphics processing units, GPUs) catalyzed the dominance of DL. Notably, the Transformer architecture,[357] with its self-attention mechanisms, underpins LLMs like ChatGPT and DeepSeek,[358,359] which demonstrate unprecedented capabilities in text generation, reasoning, and domain-specific knowledge synthesis. The evolution from symbolic to data-driven models underscores a paradigm shift toward systems that learn representations directly from data, which affords transformative potential for molecular design.[360–363]

3.2. The workflow of molecular AI methods

By leveraging AI, complex molecular interactions are modeled and interpreted, the discovery of novel molecules is accelerated,[104,364–366]

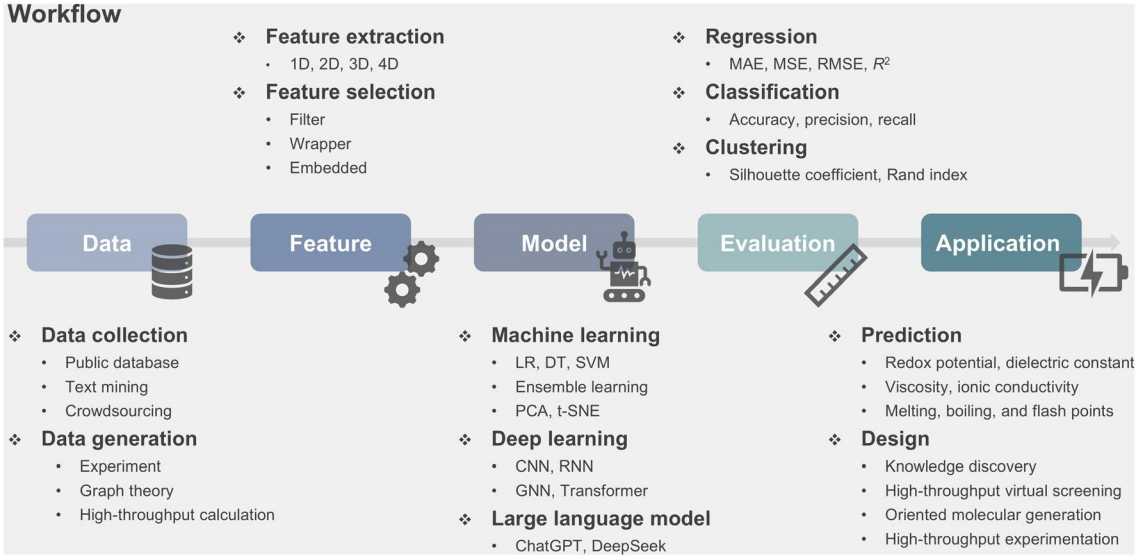

Fig. 7 The workflow of molecular AI methods.

and the development of advanced battery systems is expedited, such as electrolyte and electrode molecules with tailored properties.[367–370] The integration of AI into molecular discovery for battery research generally follows a structured workflow comprising five interconnected stages (Fig. 7).

3.2.1. Data. The foundation of AI-driven molecular discovery lies in robust data acquisition pathways, which encompass both the curation of existing datasets and the generation of new data through experimental or computational methods. Data collection primarily involves harvesting information from established repositories, scientific literature, and collaborative platforms (Table 2). Widely utilized public experimental databases for organic molecules include PubChem,[196,371] National Institute of Standards and Technology Chemistry WebBook,[372] whereas commercial indexing resources such as SciFinder,[373] which catalog molecular structures and physicochemical properties derived from empirical studies. Theoretical databases include QM7[374,375] and QM9,[376,377] which are generated using DFT calculations,[378] and molecular structure libraries such as GDB-11,[379,380] GDB-13,[381] and GDB-17,[382] which were constructed using graph-based algorithms. Efficient extraction from these repositories often employs programmatic access *via* APIs or automated web scraping tools, adhering to ethical data usage protocols and applicable licensing/terms-of-service constraints.[383–385] Beyond structured databases, unstructured textual data from research articles and patents serve as critical sources.[386–389] Text mining pipelines parse these documents into structured datasets through NLP techniques, involving tokenization, semantic feature extraction, and entity recognition to distill chemically relevant information.[390–393] Crowdsourcing further augments data collection by distributing annotation tasks to non-expert contributors, which has been successfully demonstrated in large-scale projects like ImageNet.[394] However, applying similar crowdsourcing methods in molecular science requires stringent quality control to ensure chemical accuracy.[395–398]

When existing datasets prove insufficient for specialized research objectives, targeted data generation becomes imperative. Wet-lab experimentation remains a cornerstone, particularly through high-throughput platforms that automate synthesis and characterization workflows.[436–438] For instance, coupling these systems with spectroscopic and electrochemical analysis enables parallel testing of electrolyte formulations under controlled conditions and significantly accelerates the acquisition of high-quality experimental data.[65,439,440] For structural exploration, mathematical frameworks rooted in graph theory offer systematic approaches to molecular design.[441–444] By representing atoms as nodes and bonds as edges, algorithms for graph isomorphism detection or substructure matching facilitate the combinatorial generation of chemically plausible molecules, constrained by valence rules and stability criteria.[445–448] Tools like the Surge algorithm exemplify this approach, generating isomer libraries while adhering to domain-specific constraints.[334] Meanwhile, computational chemistry bridges the gap between theoretical predictions and experimental observables. DFT calculations and MD simulations can predict critical electrolyte properties, including frontier orbital energy levels, ionic conductivity, and viscosity.[66,449]

3.2.2. Feature. A feature is a numeric representation of an aspect of raw data.[450] Reasonably selecting features can facilitate data understanding, mitigate the curse of dimensionality[451] to improve prediction performance, and enhance overall model performance.[452] Firstly, features can be ranked according to their correlation with the target variable through indicators such as the Pearson correlation coefficient[453] and mutual information (MI).[454] Subsequently, feature evaluation uses indicators such as tree-based feature importance and Shapley additive explanations (SHAP) value[455] to assess feature importance from model

Table 2 Molecular databases. The molecular databases are categorized into four types: structure, calculation, literature, and others. The literature category encompasses a broader range of sources, including scientific publications, patents, datasets, and technical reports. The database retrieval was completed as of 22 May 2025

Database	Size	Description	Ref.
Structure databases			
GDB-11	26 434 571	Up to 11 atoms (C, N, O, F)	379,380
GDB-13	99 394 177	Up to 13 atoms (C, N, O, S, Cl)	381
GDB-17	166 443 860 262	Up to 17 atoms (C, N, O, S, X)	382
Calculation databases			
AISD HOMO–LUMO	10 502 917	Based on Enamine REAL Space, including HOMO–LUMO gap	399
BatElyte	3000	Including 3000 solvents, 1.0×10^6 electrolyte formulas, and 240 000 calculation data	400
GDB-9-Ex	96 732	Based on GDB-9, ultraviolet-visible spectrophotometry (UV-Vis) absorption spectra	401
ISO17	129	Each molecule contains 5000 conformational geometries	402
Materials project (MPcules)	577 813	Small molecules, including electrochemical, thermodynamic, and more	403
MD trajectories of $C_7O_2H_{10}$	113	MD trajectories of $C_7O_2H_{10}$ isomers	402
MD17	8 Mols. 3611 115 Conf.	MD conformations	404,405
MD22	7 Mols. 223 422 Conf.	MD conformations	404,406
OMol25	>83 million	Up to 350 atoms, with DFT single-point calculations	407
ORNL_AISD-Ex	10 502 904	Including UV-Vis absorption spectra	408
PubChemQC PM6	221 190 415	Based on PubChem, including geometries and electronic properties	409
QM7	7165	Based on GDB-13, including atomization energies	261,381,402
QM7b	7211	Based on QM7, including 13 additional properties	378,381,402
QM8	21 800	Based on GDB-17, including low-lying singlet–singlet vertical electronic spectra	382,402,410
QM9 (GDB-9)	133 885	Based on GDB-17, including geometric, energetic, and other properties	376,402
Literature databases			
AAT bioquest	13 273	Including boiling point, melting point, pK_a, pK_b, and water solubility	411
ChemACX	26 965 489	Commercially available substances and find pricing and commercial availability	412
ChEMBL	2496 335	Bioactive molecules with drug-like properties	413
ChemBridge	>1.3 million	High-quality screening compounds and libraries for hit identification	414
ChemDB	407 658	Including physicochemical properties and spectroscopic properties	415
Chemexper	>1 500 000	Including suppliers, infrared spectroscopy, and nuclear magnetic resonance spectra	416
ChemSpider	>128 million	Including chemical structures, properties, and vendor information	417
Compound structure database	6 million	Including structures, physicochemical properties	418
Merck index	19 998	Including physical, pharmacological, and historical information	419
NIST chemistry webBook	—	Organic compounds, along with a few small inorganic compounds	420
Organic compounds database	2483	Including basic physical properties and spectroscopic properties	421
PubChem	121 413 818	An open chemistry database, mostly contains small molecules	422
SciFinder	>279 million	Including chemical structures, regulatory information, and properties	423
ZINC-22	37 billion	Including conformations, partial atomic charges, and solvation energies	424
Other databases			
ChemDX	571 687	Including public database, experiments database, and calculations database	425
CHEMriya	55 billion	Transform hit expansion, hit-to-lead optimization, and compound evolution	426
Enamine REAL space	64.9 billion	Compounds generated are virtual but are readily accessible	427
ESOL	2874	Including aqueous solubility	428
eXplore	>4.9 trillion	Including structure, properties, and synthesizability assessment	429
Free solvation database (FreeSolv)	643	Including experimental and calculated hydration free energies in water	430
Freedom space 4.0	142 billion	Available as a synthon-based space	431
iBonD	20 000	Including heterolytic (pK_a) and homolytic (BDE) bond dissociation energies	432
Molecular universe	100 million	A battery material discovery software and service platform, and trained a navigation system powered by a battery-specific LLM	433
ULTIMATE	>150 million	Including structure, properties, and synthesizability assessment	434
WuXi galaXi	200 million	Library of novel drug-like scaffolds	435

outputs and improve model interpretability. In addition, feature evaluation can also be conducted through ablation studies that examine model performance after removing one or more variable.[456]

From the methodological perspective, feature selection includes filter, wrapper, embedded, and hybrid methods.[457] The filter methods rank features through univariate tests (*e.g.*, analysis of variance and the F-test)[458] or multivariate analysis (*e.g.*, minimum redundancy maximum relevance).[459] The methods are suitable for the initial screening of high-dimensional data, but can neglect feature interactions and essential combinations. The wrapper method uses either forward selection, where variables are gradually incorporated into larger subsets, or backward elimination, where one starts with all variables and the least useful features are progressively removed, to ultimately result in nested subsets of variables.[457] Embedding feature selection into the model training process, the embedded methods include the least absolute shrinkage and selection operator (LASSO),[460] split gain in tree-based models, and attention mechanisms.[461] The methods are practical, but interpretability depends on the underlying model and the results are model-dependent. The hybrid methods, such as Boruta,[462] balance efficiency and accuracy by using filter methods during the initial screening process and wrapper during the fine screening process, but the complexity of parameter tuning is relatively high. Other methods, such as compressed sensing (under sparsity assumptions),[463,464] and network pruning for neural networks (where hidden units act as learned feature extractors) can also effectively reduce the dimensionality of the problem.

3.2.3. Model. AI methodologies are primarily categorized by training data utilization into three core paradigms: supervised, unsupervised, and reinforcement learning.[465–467] Supervised learning employs labeled datasets to derive input–output mapping functions, exemplified by algorithms such as linear regression and SVMs.[468–470] Unsupervised learning extracts latent patterns from unlabeled data through clustering (*e.g.*, *k*-means) or dimensionality reduction techniques like principal component analysis (PCA).[471–473] Reinforcement learning is a cross-disciplinary domain that examines how an intelligent agent can learn effective behaviors through interactions with its environment, with the objective of maximizing reward signals.[474–476] Emerging variants, including semi-supervised[477–479] and self-supervised[480,481] learning, further extend these foundational frameworks. The commonly used ML and DL methods will be discussed in Sections 3.3 and 3.4 (Fig. 8).

The efficacy of ML models hinges on rigorous evaluation frameworks involving partitioned datasets. The training set facilitates parameter optimization, while the validation set monitors generalization performance during hyperparameter tuning, preventing premature decisions on model architecture.[482–484] The test set, reserved exclusively for final evaluation, provides an unbiased estimate of model performance on unseen data.[485,486] This tripartite division supports model generalization and mitigates the risk of overfitting, where models excessively adapt to training data idiosyncrasies.[487,488] Conversely, insufficient model complexity or inadequate training often induces underfitting when models fail to capture fundamental data patterns.[489] While underfitting can be addressed by enhancing model capacity or

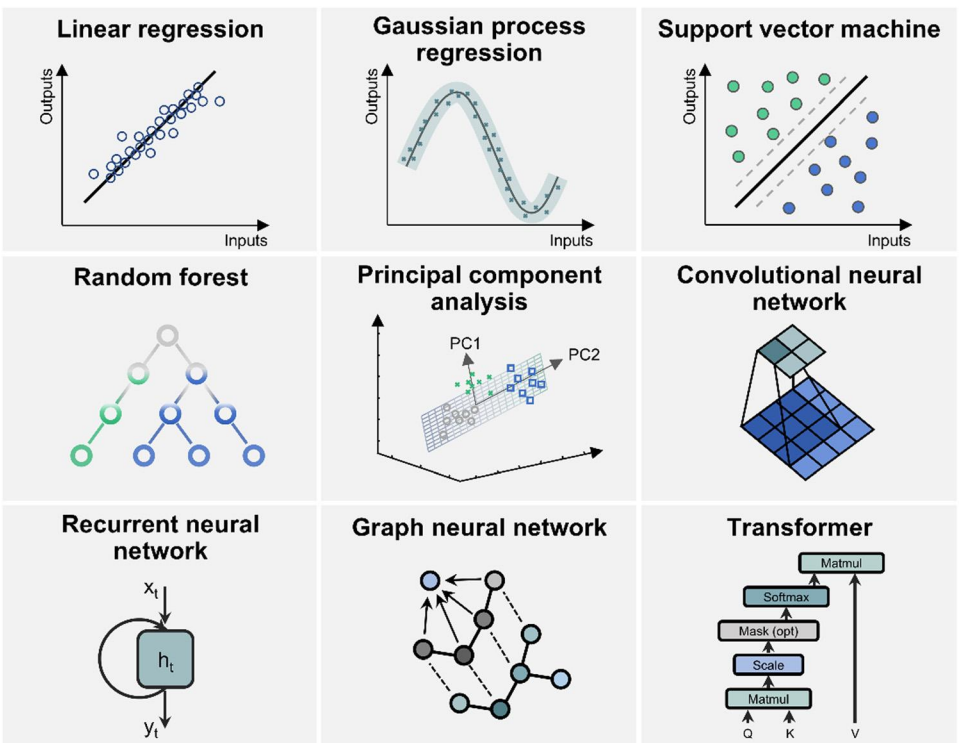

Fig. 8 The typical ML and DL methods.

extending training duration, overfitting presents an intrinsic challenge requiring systematic countermeasures.[490] Regularization techniques, such as L2 penalty terms and dropout mechanisms,[491–493] impose constraints on parameter magnitudes or network connectivity, thereby discouraging over-specialization to training noise. Central to model evaluation are the dual concepts of robustness and generalization.[494,495] Robustness quantifies the resilience of the model to perturbations in input data, ensuring stable performance under varying noise levels or adversarial conditions.[496–498] Generalization, meanwhile, reflects the model's capacity to extrapolate learned patterns to unseen data, a property directly influenced by the bias-variance tradeoff.[499,500] Striking an optimal balance between these competing factors remains a cornerstone of ML theory, as overly simplistic models can overlook critical data features (high bias), while excessively complex ones may amplify stochastic variations (high variance).[501–503]

3.2.4. Metrics

Regression performance metrics. For regression tasks, prediction performance improves when the predicted values closely match the true values. Several standard metrics quantitatively evaluate regression performance.[504–506] For instance, the mean absolute error (MAE) measures the average absolute difference between predicted and true values.

$$\mathrm{MAE} = \frac{1}{m}\sum_{i=1}^{m}|f_i - y_i| \tag{4}$$

where m is the number of samples, f_i the predicted value of the i^{th} sample, and y_i the true value of the i^{th} sample. The mean squared error (MSE) calculates the average squared difference, emphasizing large errors.

$$\mathrm{MSE} = \frac{1}{m}\sum_{i=1}^{m}(f_i - y_i)^2 \tag{5}$$

The root mean squared error (RMSE) represents the square root of MSE, maintaining the original units.

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(f_i - y_i)^2} \tag{6}$$

The coefficient of determination (R^2) indicates the proportion of variance in the dependent variable explained by the model.

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(f_i - y_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2} \tag{7}$$

where $\bar{y}$ is the mean of the true targets. Higher values of R^2, particularly those approaching 1, indicate that a greater proportion of the variance in the response variable is accounted for by the model. Note that R^2 can be negative on evaluation data if the model underperforms the mean predictor. Lower MAE, MSE, and RMSE values indicate better accuracy, while higher R^2 values reflect stronger predictive capability.

Classification performance metrics. In binary classification tasks, each sample is assigned one of two possible outcomes for both the predicted and true values.[507] As a result, four distinct combinations of predicted and actual outcomes are defined for each instance (Table 3). TP is the number of true positive samples, FN the number of false negative samples, TN the number of true negative samples, and FP the number of false positive samples.

Table 3 Four possible combinations of predicted and true values

True value	Predicted value	Symbol
Positive	Positive	True positive (TP)
Positive	Negative	False negative (FN)
Negative	Negative	True negative (TN)
Negative	Positive	False positive (FP)

To quantitatively assess model performance in binary classification, several standard metrics are employed. Accuracy is defined as the proportion of correctly classified samples (both TPs and TNs) among all samples:

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN}} \tag{8}$$

Precision is defined as the proportion of TP predictions among all predicted positive samples:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{9}$$

Recall is defined as the proportion of TP samples correctly identified among all actual positive samples:

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{10}$$

Additionally, the F1 score is introduced to capture a balance between Precision and Recall by representing their harmonic mean:

$$\mathrm{F1} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{11}$$

Furthermore, the receiver operating characteristic (ROC) curve is widely employed as a comprehensive evaluation tool for assessing classifier performance at different decision thresholds.[508,509] The ROC curve is constructed by plotting the false positive rate (FPR) on the horizontal axis and the true positive rate (TPR) on the vertical axis. The definitions of FPR and TPR are as follows:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{12}$$

$$\mathrm{TPR} = \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{13}$$

The area under the curve (AUC) is used as a quantitative indicator, and its value corresponds to the area enclosed by the ROC curve and the coordinate axes. AUC values range from 0 to 1. An AUC of 1 signifies a perfect classifier, while an AUC of 0.5

indicates equivalence to random guessing. The mathematical expression for AUC is presented as follows:

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}\ \mathrm{dFPR} \tag{14}$$

For multi-classification problems, log loss is employed to quantify the discrepancy between predicted probabilities and actual class labels.[510]

$$\log(\mathrm{loss}) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{c} q_{ij}\log p_{ij} \tag{15}$$

where m is the number of samples, c the number of categories, q_{ij} the indicator variable ($q_{ij} = 1$ when the i^{th} sample is of class j, otherwise $q_{ij} = 0$) and p_{ij} the probability of predicting the i^{th} sample as the j^{th} class ($0 \le p_{ij} \le 1$). Additionally, there are other metrics, including confusion matrix,[511,512] macro-average,[513] micro-average,[514] and weighted average.[515]

Clustering performance metrics. For clustering tasks, overall clustering quality is considered favorable when distances among samples belonging to the same category remain small, while distances among samples from different categories remain large. The evaluation metrics for clustering problems are divided into two types, depending on whether sample labels are provided.

When the samples are unlabeled, clustering metrics include the Silhouette Coefficient[449] and Davies–Bouldin index (DBI).[516,517] The Silhouette Coefficient is calculated to assess both within-cluster cohesion and between-cluster separation, and is defined as follows:

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}} \tag{16}$$

where s_i is the Silhouette Coefficient of data point i($-1 \le s_i \le 1$), b_i the minimum average dissimilarity between i and all points in any other cluster, and a_i the average dissimilarity between i and the other points in the same cluster. A negative s_i indicates that the cluster assignment of data point i can reduce cohesion and separation, whereas a value close to 1 suggests that i is more suitably placed in the assigned cluster.

The Davies–Bouldin index is a metric used to evaluate the quality of clustering algorithms by quantifying both the compactness of individual clusters and the separation between distinct clusters.

$$\mathrm{DBI} = \frac{1}{N}\sum_{i=1}^{N}\max_{j\neq i}\frac{S_i + S_j}{M_{i,j}} \tag{17}$$

where N is the number of clusters, S_i the within-cluster scatter of cluster i, and $M_{i,j}$ the distance between the centroids of clusters i and j. A lower DBI value indicates a better clustering performance.

When the samples are labeled, clustering metrics include the Rand index (RI), MI, and purity.[518,519] RI is employed as an external validation metric for assessing the consistency between a clustering result and a reference partition. Pairs of samples are examined to determine whether both clustering assignments and true labels coincide. The definition of RI is as follows:

$$\mathrm{RI} = \frac{\alpha + \beta}{\alpha + \beta + \gamma + \delta} \tag{18}$$

where α is the number of pairs assigned to the same cluster and sharing the same label, β the number of pairs assigned to different clusters and having different labels, γ the number of pairs assigned to the same cluster but having different labels, and δ the number of pairs assigned to different clusters but sharing the same labels. The value of RI ranges from 0 to 1, where higher values indicate stronger agreement between the clustering result and the reference partition.

MI measures the degree of information sharing between the clustering results and the true labels through information entropy. MI is defined as follows:

$$I(\mathbf{U};\mathbf{V}) = \sum_{i=1}^{|\mathbf{U}|}\sum_{j=1}^{|\mathbf{V}|} P(i,j)\log\frac{P(i,j)}{P_{\mathbf{U}}(i)P_{\mathbf{V}}(j)} \tag{19}$$

where $\mathbf{U}$ is the set of ground-truth classes, $\mathbf{V}$ the set of clusters produced by the clustering algorithm, $P(i, j)$ is the probability that a sample belongs to category i in $\mathbf{U}$ and cluster j in $\mathbf{V}$, $P_{\mathbf{U}}(i)$ the probability of belonging to category i in $\mathbf{U}$, and $P_{\mathbf{V}}(j)$ the probability of belonging to cluster j in $\mathbf{V}$. A higher MI value indicates a stronger alignment between the clustering result and the reference partition.

Purity calculates the proportion of categories to which most samples belong in each cluster. With $\boldsymbol{\Omega} = \{\omega_1, \omega_2, \ldots, \omega_I\}$ as a set of clusters and $\mathbf{K} = \{k_1, k_2, \ldots, k_J\}$ as a set of classes, the definition of purity is as follows:

$$\mathrm{Purity}(\boldsymbol{\Omega}, \mathbf{K}) = \frac{1}{m}\sum_i \max_j \left|\omega_i \cap k_j\right| \tag{20}$$

where m the total number of samples, $|\omega_i \cap k_j|$ is the number of samples that are simultaneously in cluster ω_i and class k_j. Higher purity indicates a better match between the clustering results and the true labels.

3.3. Molecular machine learning models

3.3.1. Linear regression. Linear regression is a foundational supervised learning model that establishes a linear relationship between an input feature matrix $\mathbf{X}$ (*e.g.*, molecular descriptors) and a target vector $\mathbf{y}$ (*e.g.*, physicochemical properties). The model parameters are optimized by minimizing a loss function that quantifies the discrepancy between predicted and observed values. The objective function is formulated as:

$$\min_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda f(\mathbf{w}) \tag{21}$$

where $\mathbf{w}$ is the parameter vector that parameterizes the linear relationship between $\mathbf{X}$ and $\mathbf{y}$, and $\lambda f(\mathbf{w})$ the regularization term, which is added to prevent overfitting. When using L1 (LASSO) regularization, $f(\mathbf{w}) = \|\mathbf{w}\|_1 = |\mathbf{w}_1| + |\mathbf{w}_2| + \cdots + |\mathbf{w}_d|$. When using L2 (Ridge) regularization, $f(\mathbf{w}) = \|\mathbf{w}\|_2^2 = \mathbf{w}_1^2 + \mathbf{w}_2^2 + \cdots + \mathbf{w}_d^2$.

The optimal parameter $\mathbf{w}_{\mathrm{opt}}$ can be obtained by calculating the analytical solution $\mathbf{w} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$ when $\lambda = 0$ and $\mathbf{X}^{\mathrm{T}}\mathbf{X}$

is invertible. For Ridge regression, $\mathbf{w} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$, whereas LASSO has no closed-form solution and is typically solved *via* coordinate descent or related algorithms. An alternative is to use (stochastic) gradient descent, which updates $\mathbf{w}$ in each iteration as follows:

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta\nabla_{\mathbf{w}}\hat{\varepsilon}(\mathbf{w})\big|_{\mathbf{w}=\mathbf{w}^{(t)}} \tag{22}$$

where $\mathbf{w}^{(t)}$ is the weight parameter in number t iteration, η the learning rate, $\nabla_{\mathbf{w}} \cdot |_{\mathbf{w}=\mathbf{w}^{(t)}}$ the gradient with respect to $\mathbf{w} = \mathbf{w}^{(t)}$, and $\hat{\varepsilon}(\mathbf{w})$ the loss function, which is $\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda f(\mathbf{w})$.

Utilizing nonlinear basis functions, linear regression can possess a nonlinear hypothesis space. Commonly employed basis functions encompass polynomial and radial basis functions. Furthermore, generalized linear models[520] extend this framework by linking the linear predictor $\mathbf{X}\mathbf{w}$ to the response variable through a link function (*e.g.*, logit for classification, exponential for Poisson regression), broadening its applicability to diverse chemical prediction tasks.

3.3.2. Gaussian process regression. Gaussian process regression (GPR) is a non-parametric modeling technique that employs Gaussian process priors for the regression analysis of data. The method can estimate hyperparameters, which control the form of the Gaussian process, through either a maximum likelihood or Bayesian approach. GPR can not only predict the mean of the target value, but also quantify the predictive distribution (mean and variance), enabling predictive/credible intervals for the prediction.

In GPR, a covariance function is required, such as the following covariance function:[521]

$$C\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) = v_0 \mathrm{e}^{-\frac{1}{2}\sum_{l=1}^{d}\mathbf{w}_l\left(\mathbf{x}_l^{(i)} - \mathbf{x}_l^{(j)}\right)^2} + a_0 + a_1\sum_{l=1}^{d}\mathbf{x}_l^{(i)}\mathbf{x}_l^{(j)} + v_1\delta(i,j) \tag{23}$$

where $\mathbf{x}^{(i)}$ is the ith training sample, v_0 the variable that determines the overall scale of local correlations, d the dimensionality of the input, $\mathbf{w}_l$ the parameter that enables a different distance measure for each input dimension, a_0 the bias term, a_1 the variable controlling the scale of linear contributions to the covariance, and $v_1\delta(i,j)$ the noise term.

Based on the covariance function, the mean and variance of a Gaussian distribution can be predicted using the following equations:[521]

$$\hat{y}(\mathbf{x}) = \mathbf{k}^{\mathrm{T}}(\mathbf{x})\mathbf{K}^{-1}\mathbf{t} \tag{24}$$

$$\sigma_{\hat{y}}^2(\mathbf{x}) = C(\mathbf{x},\mathbf{x}) - \mathbf{k}^{\mathrm{T}}(\mathbf{x})\mathbf{K}^{-1}\mathbf{k}(\mathbf{x}) \tag{25}$$

where $\hat{y}(\mathbf{x})$ is the mean, $\mathbf{k}(\mathbf{x})$ the vector defined as $(C(\mathbf{x},\mathbf{x}^{(1)}),\ldots, C(\mathbf{x},\mathbf{x}^{(n)}))^{\mathrm{T}}$, $\mathbf{K}$ the covariance matrix for the training cases, with elements $\mathbf{K}_{ij} = C(\mathbf{x}^{(i)},\mathbf{x}^{(j)})$, $\mathbf{t}$ the targets defined as $\mathbf{t} = (t^{(1)}, \ldots, t^{(n)})^{\mathrm{T}}$, and $\sigma_{\hat{y}}^2(\mathbf{x})$ the variance.

3.3.3. Support vector machine. Initially developed for addressing binary classification tasks, where samples are categorized into two groups (+1 or −1), the SVM constructs a separating hyperplane that maximizes the geometric margin between the two classes. The hard-margin primal optimization can be written as:

$$\min_{\mathbf{w},b}\frac{1}{2}\|\mathbf{w}\|_2^2 \tag{26}$$

$$\text{s.t. } y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1, \quad 1 \leq i \leq n$$

where $\mathbf{w}$ is the normal (parameter) vector of the hyperplane, b the intercept, y_i the label of the i^{th} sample point, which can take values of 1 or −1, $\mathbf{x}_i$ is the i^{th} input vector, and $\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b$ the classification score for $\mathbf{x}_i$.

To handle nonlinearity, the kernel method maps inputs to a higher-dimensional feature space *via* a feature map $\phi(\cdot)$ and replaces inner products with a kernel function:[522,523]

$$k(\mathbf{x}_1,\mathbf{x}_2) = \phi(\mathbf{x}_1)\cdot\phi(\mathbf{x}_2) \tag{27}$$

where $k(\mathbf{x}_1,\mathbf{x}_2)$ is the kernel function and $\phi(\cdot)$ the basis function.

Furthermore, the soft margin SVM was introduced, making the model suitable for handling inseparable samples. The soft-margin SVM introduced slack variables can be formulated as follows:[524]

$$\min_{\mathbf{w},b,\xi}\frac{1}{2}\|\mathbf{w}\|_2{}^2 + \zeta\sum_{i=1}^{n}\xi_i \tag{28}$$

$$\text{s.t. } y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0, \quad 1 \leq i \leq n$$

where ζ the constant that governs the trade-off between margin width and misclassification, and ξ_i the nonnegative slack variables. A training vector is misclassified if $\xi_i > 1$ and is within the margin if $0 < \xi_i < 1$.

For regression tasks, the corresponding technique is support vector regression (SVR) with an ε-insensitive loss, which introduces two nonnegative slack variables per sample to allow deviations beyond the ε-tube:[525]

$$\min_{\mathbf{w},b,\xi,\xi^*}\frac{1}{2}\|\mathbf{w}\|_2^2 + \zeta\sum_{i=1}^{n}\left(\xi_i + \xi_i^*\right) \tag{29}$$

$$\text{s.t. } (\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) - y_i \leq \varepsilon + \xi_i$$
$$y_i - (\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \leq \varepsilon + \xi_i^*,$$
$$\xi_i \geq 0,\ \xi_i^* \geq 0,\ 1 \leq i \leq n$$

where ζ balances flatness and violations, ε sets the tube width, ξ_i and ξ_i^* the nonnegative slack variables.

3.3.4. Ensemble learning. Ensemble learning is employed to enhance overall model performance through the integration of diverse models. The ensemble learning framework can be represented by the following equation:

$$h_{\mathrm{ens}}(\mathbf{x}) = \mathrm{combine}\left(h_{D_n^1}(\mathbf{x}), h_{D_n^2}(\mathbf{x}), \ldots, h_{D_n^m}(\mathbf{x})\right) \tag{30}$$

where $h_{\mathrm{ens}}(\mathbf{x})$ is the ensemble model derived from sample $\mathbf{x}$, combine$(\cdot)$ the aggregation operator (*e.g.*, averaging or majority vote), and $h_{D_n^i}(\mathbf{x})$ the base learner derived from the i^{th} subsample extracted from the original sample D_n. Base learners may be

homogeneous (as is common in Bagging and Boosting) or heterogeneous (as is typical in Stacking).

The performance advantage of ensemble learning depends on the balance between accuracy and diversity among base learners. Error-ambiguity decomposition is proposed to describe the relationship:

$$E = \bar{E} - \bar{A} \tag{31}$$

where E is the ensemble generalization error, $\bar{E}$ the average generalization error of the base learners, and $\bar{A}$ the average ambiguity among the base learners. The ensemble model can significantly outperform any single base learner when the base learners are both highly accurate and sufficiently diverse in their predictions.

Pivotal methods, such as Bagging and Boosting, are included in ensemble learning and combine outputs from several base learners to strengthen predictive capability. In the Bagging, bootstrap samples are drawn from the original dataset, and base learners are trained on these subsamples. The learners are then aggregated to reduce estimator variance.[526] As a representative Bagging algorithm, the RF method constructs multiple decision trees using both bootstrap samples and random feature subsets at each split, and concludes with a majority vote for classification or averaging for regression.[527,528] Boosting is constructed by sequentially adding weak learners and merging them into a robust final model. Unlike Bagging, the Boosting method adjusts sample weights according to the performance of the preceding weak learner, which is required to perform marginally better than random guessing. Prominent Boosting algorithms include AdaBoost, Gradient Boosting, and XGBoost.[529,530]

Ensemble learning, as a model combination technique, also incorporates averaging, voting, learning, and other approaches. In the domain of DL, ensemble concepts have been broadly applied in scenarios like model averaging and knowledge distillation, forming a crucial technique for performance improvements in complex tasks.[523,531]

3.3.5. k-means clustering. As a frequently employed unsupervised learning method, the k-means algorithm partitions sample data into k clusters by initially designating k center points $\{\mu^{(1)},\ldots,\mu^{(k)}\}$ and subsequently executing the following two steps iteratively until convergence. Firstly, allocate each training sample to the cluster i denoted by the closest center point $\mu^{(i)}$. Secondly, update each center point $\mu^{(i)}$ to the arithmetic mean of all training samples within cluster i.[523]

The application of the k-means algorithm necessitates adherence to the following conditions. There are a few attributes per instance to avoid expensive computation. A sufficient number of samples must be available to mitigate the curse of dimensionality.[532] Given that the feature scale can dynamically impact clustering outcomes, it is imperative to preprocess the data using the following feature normalization formula:

$$\hat{x}_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{32}$$

where $\hat{x}_{ij}$ is the normalized value of feature j for sample i, μ_j the mean of sample feature j, $\mu_j = \frac{1}{m}\sum_{i=1}^{m} x_{ij}$ (m represents the number of samples), and σ_j the standard deviation of sample feature j, defined as $\sigma_j = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_{ij} - \mu_j)^2}$. Finally, select a suitable positive integer k, as the model may overfit if k is too large and underfit if k is too small.

However, the k-means method is subject to several limitations.[532] It assumes roughly spherical, similarly sized clusters and is sensitive to initialization and outliers. Performance may degrade in high-dimensional spaces due to distance concentration, underscoring the importance of feature engineering and dimensionality reduction. To address these issues and adapt the method to various scenarios, new variants and related approaches have been proposed, including k-means++ initialization, mini-batch k-means, and k-medoids.[533]

3.3.6. Principal component analysis. PCA is widely recognized as a fundamental technique for dimensionality reduction in data-driven research. It projects high-dimensional data onto a lower-dimensional subspace while preserving key information. Principal components are formed by linearly recombining sample features in a way that ensures mutual orthogonality, therefore retaining a substantial amount of data variance with minimal memory usage. The property is especially useful for mitigating the effects of redundant features and reducing computational requirements.[534]

Two classical derivations exist, namely minimum reconstruction error and maximum separability. The former is adopted in this section. The columns of the data matrix are centered to obtain $\mathbf{X}$, and the sample covariance matrix is then computed to capture relationships among the features: $\mathbf{S} = \mathbf{X}^{\mathrm{T}}\mathbf{X}/(m-1)$. Principal components can be determined either through the eigendecomposition of $\mathbf{S}$ or through the singular value decompositionrs of $\mathbf{X}$. Both methods identify the eigenvectors (or singular vectors) with the largest eigenvalues, which correspond to directions of maximal variance in the feature space. Mathematically, PCA can be formulated as an optimization problem under the minimum-reconstruction-error principle:[523]

$$\mathbf{d}^* = \underset{\mathbf{d}}{\operatorname{argmin}} \left|\left|\mathbf{X} - \mathbf{X}\mathbf{d}\mathbf{d}^{\mathrm{T}}\right|\right|_F^2 \tag{33}$$

$$\text{s.t. } \mathbf{d}^{\mathrm{T}}\mathbf{d} = 1$$

where $\mathbf{d}^*$ is the first principal component (optimal $\mathbf{d}$), which corresponds to the eigenvector associated with the maximum eigenvalue of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, $\|\cdot\|_F^2$ the squared Frobenius norm, and $\mathbf{X}$ the centered data matrix with m rows and n columns, where m represents the number of samples and n represents the number of features.

Subsequent principal components are determined by extracting additional eigenvectors associated with the next largest eigenvalues, ensuring orthogonality to previously identified components. The number of principal components retained can be decided based on variance thresholds or other criteria, depending on the specific requirements of the analysis. PCA is frequently applied for tasks such as data visualization, noise

reduction, and feature selection, contributing to more efficient learning and improved interpretability of complex datasets.

3.3.7. t-Distributed stochastic neighbor embedding. t-Distributed stochastic neighbor embedding (t-SNE) is employed to visualize high-dimensional data by projecting them onto 2D or 3D spaces. A student-t distribution with one degree of freedom is employed to alleviate the crowding problem and to improve optimization stability.[535,536] During t-SNE, it is essential to calculate the symmetrized joint probabilities p_{ij} in the high-dimensional space and the joint probabilities q_{ij} in the low-dimensional space using the following formulations:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2m} \tag{34}$$

$$q_{ij} = \frac{\left(1 + \left\|y_i - y_j\right\|_2^2\right)^{-1}}{\sum_{k \neq l}\left(1 + \left\|y_k - y_l\right\|_2^2\right)^{-1}} \tag{35}$$

where m is the number of samples, $p_{j|i}$ is the conditional probability signifying the similarity between data points x_j and x_i. $p_{j|i}$ can be calculated as

$$p_{j|i} = \frac{e^{-\frac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2^2}{2\sigma_i^2}}}{\sum_{k \neq i} e^{-\frac{\left\|\mathbf{x}_i - \mathbf{x}_k\right\|_2^2}{2\sigma_i^2}}} \tag{36}$$

where σ_i is the standard deviation of the Gaussian centered on data point $\mathbf{x}_i$. The t-SNE algorithm strives to minimize the Kullback-Leibler divergence between the two joint probability distributions (P and Q):

$$\mathrm{KL}(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{37}$$

In t-SNE, perplexity is regarded as a key hyperparameter that controls the size of the local neighborhood. Low perplexity values emphasize local structural details, whereas higher perplexity values preserve broader global relationships. A typical recommended perplexity range is from 5 to 50. Subsequent improved algorithms, such as uniform manifold approximation and projection (UMAP), were proposed to enhance global structure preservation and computational efficiency by incorporating topological constraints and more flexible similarity metrics.[537]

3.4. Molecular deep learning models

3.4.1. Artificial neural networks. ANNs are computational models inspired by the structure and functionality of the human brain, in which neurons are highly interconnected and capable of rapidly transmitting information. In ANNs, artificial neurons serve as the fundamental units, each comprising three primary components, including weights, biases (also referred to as thresholds), and activation functions. The computational operation of a single neuron is expressed as

$$y = \mathrm{neuron}(\mathbf{x}; \mathbf{w}, b) = f(\mathbf{w}^{\mathrm{T}}\mathbf{x} + b) \tag{38}$$

where y is the output of a neuron, $\mathbf{x}$ the input vector, b the bias of a neuron, and $f(\cdot)$ the activation function. Common activation functions include the rectified linear unit (ReLU), sigmoid, and hyperbolic tangent, each introducing nonlinearity that enables the network to capture complex relationships.

The most basic ANN is the feedforward neural network (FNN). It consists of an input layer, one or more hidden layers, and an output layer. During computation, information is propagated forward through the network, and the parameters $\mathbf{w}$ and b are updated in the neurons in reverse through the backpropagation algorithm.

The nonlinearity brought by activation functions and the combination of different functions in ANN bring powerful fitting ability to ANN. The universal approximation theorem states that if a FNN has a linear output layer and at least one hidden layer with activation functions, as long as a sufficient number of hidden units are given to the network, it can approximate any Borel measurable function from a finite-dimensional space to another finite-dimensional space with arbitrary accuracy.[538,539] This foundational property establishes the theoretical basis for the wide applicability of ANNs across a broad range of learning tasks.

3.4.2. Convolutional neural networks. CNNs refer to a class of neural networks that employ convolution operations rather than general matrix multiplications in at least one layer.[523] CNNs are specifically designed to analyze data characterized by a grid-like structure. A typical CNN architecture comprises convolutional layers followed by pooling layers; pooling reduces spatial dimensionality, whereas convolution may preserve or change it depending on stride and padding, while preserving key features.

Convolutional layers involve three primary principles: sparse connectivity, parameter sharing, and translation equivariance of the linear convolutional operator. Sparse connectivity ensures that each output neuron is linked to only a localized area of the input, reducing computational complexity. Parameter sharing applies the same filters across different input regions, decreasing memory usage. Translation equivariance preserves relative spatial relationships, enabling similar patterns to be recognized under shifts of the input; pooling then introduces partial translation invariance.

During convolution, multiple learnable filters slide over the input to produce feature maps, each containing linear responses that are then passed through an activation function to inject nonlinearity. When processing a 2D image $\mathbf{I}$, a typical convolution at position (i,j) is expressed as

$$S(i,j) = (\mathbf{I} * \mathbf{K}_{\mathrm{r}})(i,j) = \sum_m \sum_n \mathbf{I}(m,n)\mathbf{K}_{\mathrm{r}}(i-m, j-n) \tag{39}$$

where $\mathbf{K}_{\mathrm{r}}$ is the 2D kernel. Pooling layers then aggregate local statistical information from these feature maps to reduce their size, introducing partial invariance to small spatial shifts in the input.

Beyond image recognition, CNNs have demonstrated significant potential in molecular studies by mapping chemical structures into grid-like representations. AtomNet was introduced as the first structure-based deep CNN designed to predict the bioactivity of small molecules.[540] Chemception relies exclusively on 2D images of molecules for chemical property predictions,

matching the performance of expert-developed QSPR models.[541] ImageMol harnesses large-scale molecular image datasets by integrating an image-processing framework with extensive chemical knowledge, thereby extracting fine-grained, pixel-level features.[542] Further examples include DeepVS,[543] 2DConvNet,[287] MolMapNet,[293] and MolNexTR,[295] addressing distinct applications of CNNs in molecular discovery.

3.4.3. Recurrent neural networks. RNNs exhibit strong capabilities in handling sequential data, including applications in text comprehension and language recognition. In RNN architectures, the same weights are shared across time steps (temporal parameter sharing), thereby reducing the number of learnable parameters. The update equations can be described as follows:

$$\mathbf{s}_t = f(\mathbf{W}\mathbf{s}_{t-1} + \mathbf{U}\mathbf{x}_t + \mathbf{b}) \tag{40}$$

$$\mathbf{o}_t = \mathbf{V}\mathbf{s}_t + \mathbf{c} \tag{41}$$

where $\mathbf{s}_t$ is the state at time t (starting from the initial state $\mathbf{s}_0$), $f(\cdot)$ the activation function, $\mathbf{W}$, $\mathbf{U}$, $\mathbf{V}$ the parameter matrices, $\mathbf{x}_t$ the input at time t, $\mathbf{b}$, $\mathbf{c}$ the parameters, $\mathbf{o}_t$ the output at time t, and the standardized probability output $\hat{\mathbf{y}}_t = \mathrm{softmax}(\mathbf{o}_t)$.

There are two important variants of RNN. The long short-term memory (LSTM) network augments the recurrent unit with gates (input, forget, and output) and a cell state, which mitigates vanishing gradients and enables long-range dependency modeling.[523,544,545] Bidirectional RNN achieves the output of predicted values based on the entire input sequence by transmitting information in the forward direction from the starting point and in the backward direction from the end of the sequence.[546–548]

In molecular representation, string formats such as SMILES are amenable to modeling through RNNs. BIMODAL is introduced as a bidirectional generative RNN based on SMILES, enabling the generation of novel molecules from scratch using string representations.[549] The cRNN method is further applied to incorporate chemical conditions, allowing molecules to be generated that satisfy specified requirements.[550] Additional RNN-based approaches, including LSTM,[551] MolecularRNN,[552] QBMG,[553] M-RNN,[554] and ChemTSv2,[555] have also been developed for applications in molecular design.

3.4.4. Graph neural networks. A graph consists of nodes (vertices) and edges. Small molecules can therefore be modeled as graphs by representing atoms as nodes and chemical bonds as edges.[556] Early methods, such as DeepWalk[557] and Node2vec,[558] learn static embeddings for each node. DeepWalk estimates the similarity between node u and v using co-occurrence probabilities from truncated random-walk sequences.[557] Node2vec trades off between local and global views of the network by introducing the return parameter and the in-out parameter respectively.[558] The objective function used in the above-mentioned Node2vec work is as follows:[558]

$$\max_f \sum_{u\in V} \sum_{n_i\in N_s(u)} \log \frac{\mathrm{e}^{f(n_i)\cdot f(u)}}{\sum_{v\in V} \mathrm{e}^{f(v)\cdot f(u)}} \tag{42}$$

where $f(\cdot)$ is the mapping function from nodes to feature representations, $u \in V$ the source nodes that belong to the vertices V, $n_i \in N_s(u)$ the neighborhood nodes that belong to $N_s(u)$, which is a network neighborhood of node u generated through a neighborhood sampling strategy S, and $\sum_{v\in V} \mathrm{e}^{f(v)\cdot f(u)}$ the per-node partition function.

In contrast, GNNs learn node embeddings through end-to-end neural message passing. GNN architectures, specifically GCN,[559] GraphSAGE,[560] and graph attention networks,[561] were introduced for applications in molecular modeling. Then, the MPNN was introduced to unify various graph learning approaches under a generalized framework.[562] The forward pass has two phases, a message passing phase and a readout phase. During the message passing phase, hidden states $\mathbf{h}_v^{(t)}$ at each node v are iteratively updated through messages $\mathbf{m}_v^{(t+1)}$ derived from neighboring nodes. Formally, the phase is governed by two learnable functions: the message function M_t and the vertex update function U_t. The message from neighboring nodes is computed as

$$\mathbf{m}_v^{(t+1)} = \sum_{w\in N(v)} M_t\left(\mathbf{h}_v^{(t)}, \mathbf{h}_w^{(t)}, \mathbf{e}_{vw}\right) \tag{43}$$

where $\mathbf{m}_v^{(t+1)}$ is the message at node v at time step $t+1$, $w \in N(v)$ the neighborhood nodes that belong to $N(v)$, which denotes the neighbors of v in graph G, M_t the message function, $\mathbf{h}_v^{(t)}$ the hidden state of node v at time step t, $\mathbf{h}_w^{(t)}$ the hidden state of neighborhood node w at time step t, and $\mathbf{e}_{vw}$ the edge features between node v and neighborhood node w. The new hidden state of node v is then updated by

$$\mathbf{h}_v^{(t+1)} = U_t\left(\mathbf{h}_v^{(t)}, \mathbf{m}_v^{(t+1)}\right) \tag{44}$$

where U_t is the vertex update function. After T message passing steps, the readout phase generates a feature vector for the entire graph, by applying a permutation-invariant readout function R to the set of node representations.

$$\hat{\mathbf{y}} = R\left(\left\{\mathbf{h}_v^{(T)} \middle| v \in G\right\}\right) \tag{45}$$

where $\hat{\mathbf{y}}$ is the feature function for the whole graph G, R the readout function, and $\mathbf{h}_v^{(t)}$ the hidden state of node v at the final time step T.

Many GNNs treat molecules as 2D graphs, overlooking 3D distances and angles that underpin molecular conformations. Simply appending coordinates often renders models sensitive to translations and rotations. To address the problem, GNNs with 3D geometric information have emerged, ensuring consistent outputs despite rigid transformations.[563] For instance, the deep tensor neural network pioneered distance-based message passing,[564] and SchNet extended the concept with continuous filters capturing local atomic correlations.[565] Later models such as HIP-NN,[566] PhysNet,[567] DimeNet,[568] DimeNet++,[569] OrbNet,[570] OrbNet-Equi,[571] SphereNet,[572] ComENet,[573] and GeoGNN[112] integrate angular terms or orbital information. Additionally, SE(3)-equivariant molecular networks have been widely applied to molecular modeling due to their inherent advantages in preserving geometric symmetries, enabling efficient 3D structural learning, and maintaining consistent physical representations

under rotational/translational transformations. For example, tensor field neural networks are designed so that each layer remains locally equivariant to 3D rotations, translations, and permutations of points.[574] Because rotation equivariance is embedded in the architecture, the requirement for data augmentation is reduced, and features can be identified in arbitrary orientations. The polarizable atom interaction neural network (PAINN) has been proposed as an equivariant network and has demonstrated superior performance compared to invariant networks.[575] Vector-valued features are introduced in the embedding stage of PAINN, complementing scalar descriptors and providing a richer representation that captures complex intramolecular interactions. Furthermore, representative examples include Cormorant,[576] L0/L1Net,[577] EGNN,[578] GemNet,[579] ClofNet,[580] NequIP,[126] MACE,[581] LEFTNet,[582] SEGNO,[583] and DEGNN.[584]

3.4.5. Transformer. The Transformer has become the predominant neural architecture in DL, owing to the attention-based design and parallelizable computations. An encoder–decoder stack is built from identical layers that combine multi-head self-attention with position-wise FNNs, with each sublayer wrapped by residual connections and layer normalization for stable optimization.[461]

The core operation is scaled dot-product attention, defined as follows:

$$\mathrm{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right)\mathbf{V} \tag{46}$$

where $\mathbf{Q}$ is the query matrix, $\mathbf{K}$ the key matrix, $\mathbf{V}$ the value matrix, d_{k} the dimension of keys, $\sqrt{d_{\mathrm{k}}}$ the scaling factor, which is added to improve numerical stability, and $\mathrm{softmax}(\cdot)$ the softmax function.

To capture diverse relational patterns, attention heads are evaluated in parallel and concatenated:

$$\mathrm{MultiHead}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{Concat}(\mathrm{head}_1,\ldots,\mathrm{head}_h)\mathbf{W}^{(\mathrm{O})} \tag{47}$$

$$\mathrm{head}_i = \mathrm{Attention}(\mathbf{Q}\mathbf{W}_i^{(\mathbf{Q})},\mathbf{K}\mathbf{W}_i^{(\mathbf{K})},\mathbf{V}\mathbf{W}_i^{(\mathbf{V})}) \tag{48}$$

where $\mathbf{W}_i^{(\mathbf{Q})}$, $\mathbf{W}_i^{(\mathbf{K})}$, and $\mathbf{W}_i^{(\mathbf{V})}$ are learned linear projections for queries, keys, and values, and $\mathbf{W}^{(\mathrm{O})}$ the output projection that maps the concatenation back to the model dimension.

Because attention alone lacks positional awareness, a position vector is added to each token embedding vector before the first layer, either by fixed sinusoidal functions or by learned parameters, enabling the network to distinguish the sequence order without recurrence. Through these components, the Transformer attains efficient global context modelling, high scalability, and state-of-the-art performance across language, vision, and scientific sequence tasks.[585]

The Transformer architecture, due to its powerful attention mechanism and flexibility, has emerged as one of the most prominent DL models in molecular modeling. Specifically, the SE(3)-Transformer utilizes a specialized self-attention mechanism designed for three-dimensional point clouds and graphs.[586] Equivariance under rotations and translations is ensured by constraining the attention weights to remain invariant to such transformations. Consequently, physical consistency is preserved throughout molecular modeling tasks. Uni-Mol further extends Transformer-based molecular modeling by offering a universal 3D molecular representation learning framework.[136] It leverages an SE(3)-Transformer backbone pretrained on large-scale datasets of molecular conformations and protein pockets. Uni-Mol incorporates comprehensive 3D information directly into the representations, enabling highly accurate molecular property predictions and 3D spatial tasks. Building on these successes, several other Transformer-based architectures have been proposed, including Molecular Transformer,[587] LieTransformer,[588] 3D-Transformer,[589] Transformer-M,[590] Graphormer,[591,592] TorchMD-Net,[593] TorchMD-Net 2.0,[594] MolGPT,[595] Uni-Mol+,[596] Uni-Mol2,[597] and BAMBOO.[598]

3.5. Large language models and prospects

3.5.1. Concepts and scaling laws. LLMs are deep neural networks characterized by transformer-based architectures trained on extensive text corpora, enabling human-like language understanding and generation through self-supervised learning.[599] The term large denotes two critical dimensions: (1) massive-scale training data, typically spanning hundreds of billions to trillions of tokens (*e.g.*, DeepSeek-V3 trained on 14.8 trillion tokens), and (2) enormous parameter counts, often exceeding billions to hundreds of billions (*e.g.*, DeepSeek-V3 with 671 billion total parameters). These scales synergistically empower LLMs to capture intricate linguistic patterns, domain-specific knowledge, and cross-task generalization.

Scaling laws mathematically characterize the relationship between model performance and three fundamental variables: model parameter count (N), training data size (D), and computational resources (C). These laws are typically expressed *via* empirical power-law formulations that quantify how the loss diminishes as these factors scale. A generalized formulation can be represented as

$$L(N,D,C) = \frac{A}{N^{e_1}} + \frac{B}{D^{e_2}} + \frac{G}{C^{e_3}} \tag{49}$$

where A, B, G, e_1, e_2, and e_3 are empirically derived constants. The framework posits that increasing N, D, or C reduces loss, but their contributions are non-linear and interdependent, necessitating strategic trade-offs.

Kaplan *et al.*[600] (OpenAI team) first proposed scaling laws, prioritizing model parameters as the primary driver of performance. Building on this, Hoffmann *et al.*[600] (Google DeepMind team) demonstrated that balanced model-data scaling optimizes computational efficiency, shifting focus from pure model-centric approaches. Further advancing this trajectory, Bi *et al.*[601] (DeepSeek AI team) revealed that high-quality data reorient optimal scaling toward model expansion over data volume while formalizing hyperparameter-compute power-law relationships. Collectively, these works chart the shift in scaling strategies from parameter-centric to holistic approaches balancing data quality, computing, and architecture.

3.5.2. Basic technique. The advancement of LLMs has been largely enabled by the Transformer architecture, a neural framework employing self-attention to process sequential data

in parallel.[602] This innovation addressed the computational inefficiencies of prior architectures, enabling scalable training on massive text corpora. Early models like BERT[603] and GPT[604] demonstrated the Transformer's versatility: BERT utilized bidirectional context for masked token prediction, while GPT adopted autoregressive generation to model text sequences. These models established the pre-training paradigm, where self-supervised learning on unlabeled text, *via* objectives like next-token prediction (GPT) or masked token recovery (BERT), captured broad linguistic patterns, forming the basis for transfer learning.

To adapt pre-trained models to downstream tasks, fine-tuning emerged as a standard approach, updating model parameters using task-specific labeled data. However, its computational cost and overfitting risks prompted innovations like instruction tuning, which fine-tunes models on diverse tasks formatted with natural language instructions.[605] Instruction tuning has been shown to enhanced zero-shot generalization by by exposing models to diverse instruction formats, and reinforcement learning from human feedback further aligns model behavior with human preferences. Concurrently, supervised fine-tuning refined outputs using human-annotated data, aligning responses with safety or stylistic guidelines. These methods shifted focus from brute-force scaling to efficient specialization, balancing generalization with task-specific precision.

During inference, the phase where models generate outputs without weight updates, techniques like prompt engineering optimized input design. By embedding task descriptions into prompts, models inferred desired behaviors without retraining. This approach later evolved into chain-of-thought prompting, which elicited step-by-step reasoning to improve complex problem-solving.[606] Later, retrieval-augmented generation integrated external knowledge retrieval during inference, grounding outputs in factual data to mitigate hallucinations.[607] Collectively, these advancements prioritized efficiency and controllability, reflecting a trajectory from architectural innovation to practical deployment, where scalability, adaptability, and precision are systematically balanced.

3.5.3. Representative models. The rapid advancement of LLMs such as GPT-4o,[608] DeepSeek-V3,[359] Claude 3.5,[609] Llama 3.1,[610,611] and Qwen 2.5[612] has reshaped AI research and application landscapes (Fig. 9). OpenAI's foundational contributions began with transformer-based GPT models. GPT-2, released in 2019, had 1.5 billion parameters, whereas GPT-3, introduced in 2020, dramatically expanded to 175 billion parameters. In 2022, OpenAI released ChatGPT, a groundbreaking conversational AI that quickly gained global attention due to its remarkable conversational capabilities. Then, OpenAI introduced GPT-4o in 2024,[608] a flagship model enabling real-time multimodal reasoning across audio, vision, and text. The o1 series emphasizes test-time compute and deliberate reasoning, achieving strong results in mathematics, coding, and scientific tasks.[613]

While GPT-series models have demonstrated exceptional performance, their proprietary nature partially impedes scientific research. In contrast, DeepSeek AI, a representative open-source LLM established in 2023, achieves a significant balance between computational efficiency and model performance. The initiative pioneered the development of its mixture-of-experts architecture, which employs sparse activation mechanisms to minimize energy consumption. The 2024 DeepSeek-V2 model (236 billion parameters, 21 billion active) introduced multi-head latent attention for optimized inference.[614] Later, DeepSeek-V3 achieved breakthroughs as a leading open-source model (671 billion parameters, 37 billion active/token), outperforming other open-source models and matching leading closed-source reasoning performance with only 2.788 million H800 GPU training hours.[359] In 2025, DeepSeek-R1 integrated self-verification and extended chain-of-thought, attaining parity with advanced models like OpenAI-o1-1217 in mathematical precision.[615]

3.5.4. Prospects in battery research. LLMs exhibit considerable potential for advancing battery molecule discovery, particularly when employed in text mining to address the challenge of accelerating knowledge extraction from the exponentially growing body of scientific literature. For instance, Zheng *et al.*[616] demonstrated the effectiveness of prompt

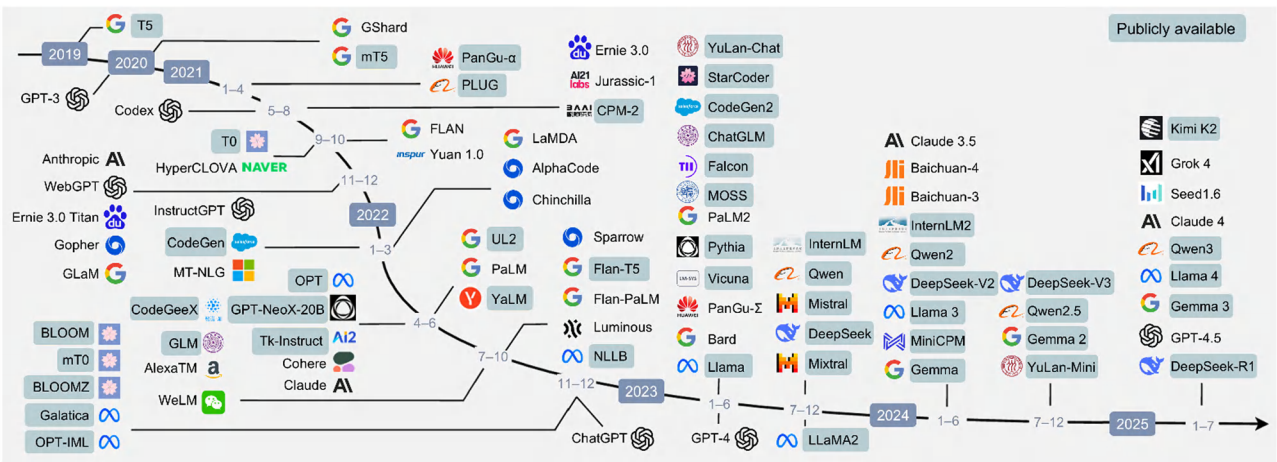

Fig. 9 The evolution of LLMs.[599] The models have been updated since July 2025. Reproduced with permission from ref. 599. Copyright 2025 The Authors.

engineering in guiding ChatGPT to automatically extract metal–organic framework (MOF) synthesis conditions from diverse publication formats and styles, resulting in 26 257 synthesis parameters linked to approximately 800 MOFs. Similarly, Na *et al.*[617] applied an LLM-driven approach for sodium-ion battery layered cathode materials, rapidly extracting 945 data points pertaining to composition, crystallinity, operating voltage, and electrode composition from 312 publications, with each paper processed in around 20 seconds. Further illustrating the efficiency of LLM-based systems, Zhao *et al.*[618] developed BatteryGPT, showcasing a platform capable of rapid literature summarization, knowledge retrieval, and question answering, thereby highlighting the powerful capacity of LLMs for information synthesis and insight generation.

While LLMs excel in text-related tasks, handling graph-structured data remains challenging. To bridge this gap, Wang *et al.*[619] introduced Graph2Token, which aligns graph tokens with LLM tokens by mapping graph elements to the model's vocabulary, thereby improving molecular prediction. Zheng *et al.*[616] proposed LLM4SD, which leverages LLMs to drive scientific discovery in molecular property prediction. By mining established information from the literature, such as molecular weight as a key indicator for solubility, and identifying patterns like the tendency of halogen-containing molecules to penetrate the blood–brain barrier, LLM4SD converts these insights into interpretable feature vectors. When integrated with interpretable models such as RFs, these features enable LLM4SD to surpass state-of-the-art benchmarks in various property prediction tasks.

Beyond text mining and property prediction, LLMs are beginning to function as AI chemists. Coscientist,[620] powered by GPT-4, exemplifies this trend by autonomously designing, planning, and executing complex experimental workflows through an integrated platform that includes internet and document searches, code execution, and experimental automation. Its capabilities were demonstrated by successfully optimizing several research tasks, including a palladium-catalyzed cross-coupling reaction, illustrating its advanced capacity for (semi)-autonomous experimental design and execution. Song *et al.*[621] reported a robot AI chemist underpinned by ChemAgents, a hierarchical multi-agent system based on the onboard Llama-3.1-70B LLM, capable of performing intricate multi-step experiments with minimal human intervention. Governed by a task manager agent that interacts with human researchers and coordinates four role-specific agents (literature reader, experiment designer, computation executor, and robot operator), this system effectively navigated a vast five-component chemical space to identify a high-performance high-entropy metal–organic catalyst for the oxygen evolution reaction.

4. Applications to molecular property prediction in rechargeable batteries

Following the comprehensive introduction of molecular representations and AI methods, accurately predicting electrolyte properties becomes essential. Although experimental characterization and classical computational simulations provide property data, these approaches are expensive, time-consuming, or lack sufficient accuracy for HTVS. In contrast, AI-based prediction methods have been widely adopted to overcome these challenges.

In this section, several key molecular properties relevant to rechargeable batteries are systematically discussed. Initially, redox potentials of electrolyte and organic electrode molecules are introduced. Subsequently, critical factors influencing molecular interactions, such as dielectric constants and donor numbers (DNs), are analyzed. Then, important characteristics affecting ionic transport, particularly viscosity and ionic conductivity, are presented. Finally, other fundamental physicochemical properties, including melting points, boiling points, and flash points, are addressed. Representative studies and key advancements in molecular property prediction using AI models will also be highlighted.

4.1. Redox potential

The redox potential is a fundamental thermodynamic property that describes electrochemical stability. Specifically, the redox potential of organic molecules plays an essential role in designing battery electrolytes and organic electrodes. The redox potential of electrolyte molecules is a critical determinant of battery performance. For instance, increasing and reducing the redox potential of catholytes and anolytes, respectively, is beneficial for increasing the energy density of redox flow batteries. To enable rapid screening of candidate molecules, Jia *et al.*[622] proposed a graph-based ML approach for predicting redox potentials. GNN models were developed to correlate molecular properties and descriptors, and an MAE of 5.6 and 7.2 kcal mol^{-1} was achieved for predicting reduction and oxidation potentials, respectively. To narrow the gap with DFT-based methods, Hruska *et al.*[623] proposed ML models to correct computational errors in redox potential calculations obtained through implicit and explicit solvent models (Fig. 10a). Physically inspired features were employed, including solute molecular properties and solvent-related features. After applying the ML correction, the MAE for implicit and explicit solvent model data was reduced from 0.76 to 0.44 V and from 0.64 to below 0.24 V, respectively. To accelerate traditional simulations, an MLMD-based method was devised by Wang *et al.*,[624] enabling automated prediction of redox potentials with high accuracy.[625–628] Concurrent learning workflows were combined with free energy calculation techniques to construct ML potentials for efficient and precise free energy estimation. Nanosecond-scale simulations with first-principles accuracy were performed using MLMD, and an automated workflow incorporating the hybrid HSE06 functional was employed for potential development. An MAE of 0.33 V was achieved in redox potential predictions, marking a significant step toward high-throughput computational screening of novel electrolytes.

To better align with experimental measurements, Gao *et al.*[629] constructed an experimental database encompassing over 500 redox potentials of organic redox-active molecules measured in aqueous and/or organic solvents (Fig. 10b). Data were collected from hundreds of published papers, recorded molecular structures

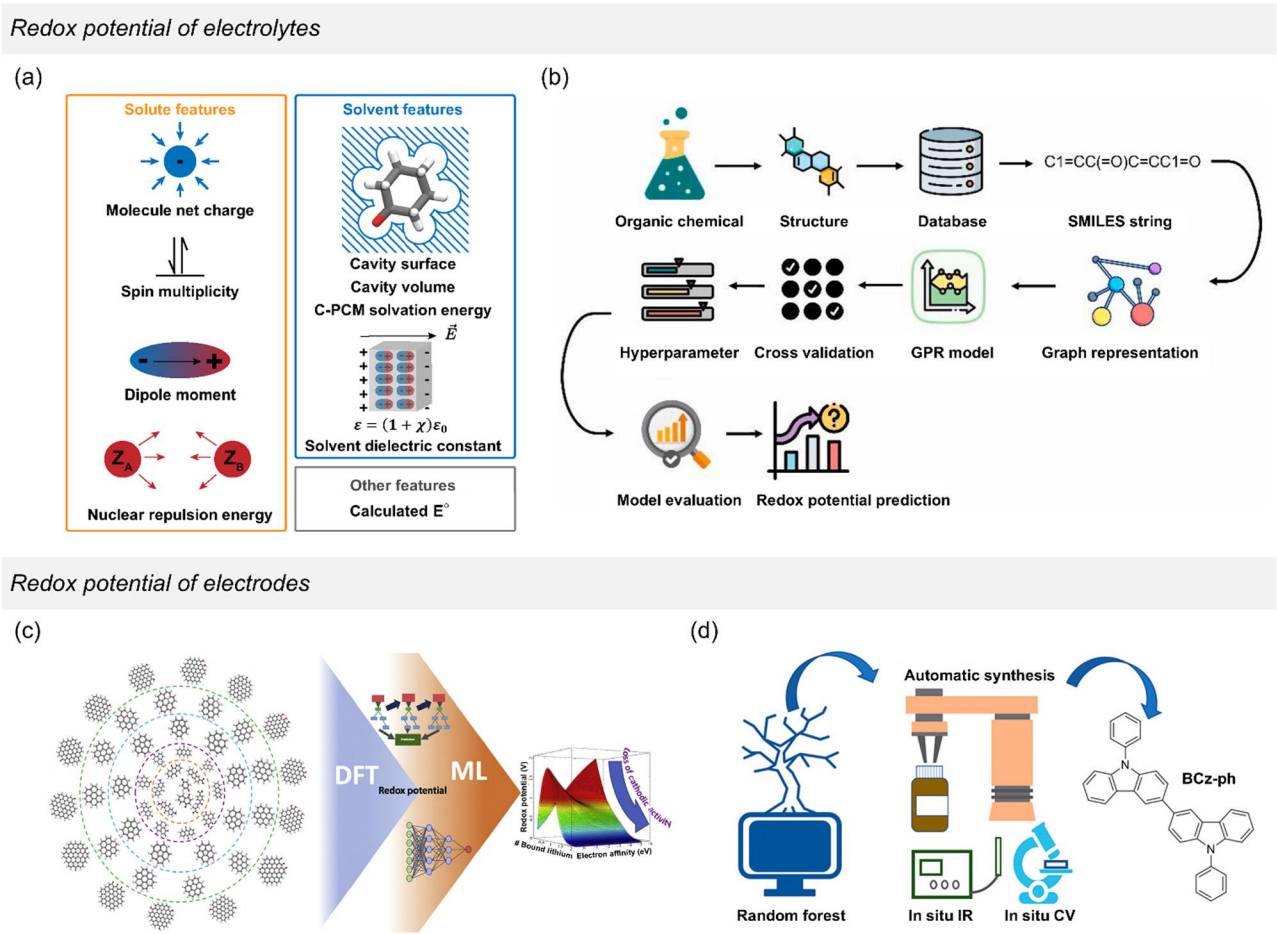

Fig. 10 The prediction of redox potentials for electrolyte and organic electrode molecules. (a) The ML-corrected solvent modeling using implicit/explicit solvent features to refine redox potential calculations.[623] Reproduced with permission from ref. 623. Copyright 2022 American Chemical Society. (b) The multi-stage ML workflow for predicting redox-active molecule performance in organic redox flow batteries.[629] Reproduced with permission from ref. 629. Copyright 2025 Elsevier Inc. (c) The HTVS combining DFT and ML for novel organic electrode molecular design.[630] Reproduced with permission from ref. 630. Copyright 2020 Elsevier Inc. (d) The ML-driven material discovery pipeline targeting high-voltage organic compounds for rechargeable battery applications.[631] Reproduced with permission from ref. 631. Copyright 2021 American Chemical Society.

using SMILES, and computed molecular descriptors using RDKit. To control for potential influences from solvent types and solution pH values, the dataset was divided into three subsets (aqueous, alkaline, and organic solvents), with resulting model performances on test sets of 0.053, 0.085, and 0.118 V, respectively. These methods provide valuable tools for the rapid exploration of high-performance redox flow batteries.

In Li battery systems, the decomposition of electrolytes at the electrode surface results in significant capacity loss and deteriorated cycling performance. Ideally, the decomposition can be suppressed by forming a stable solid electrolyte interphase (SEI) on the electrode surface.[632] The redox potential is among the fundamental properties of electrolyte film-forming additives, as additives usually preferentially participate in interfacial redox reactions compared to the bulk electrolyte, thus protecting the electrolyte from further consumption and mitigating capacity degradation. Okamoto *et al.*[122] computationally investigated the redox potentials of 149 LIB electrolyte additives using *ab initio* calculations. Twenty-two molecular descriptors were constructed based on the constituent elements and coordination numbers within the molecules. By employing Gaussian kernel ridge regression and gradient boosting regression (GBR), the state-of-the-art model achieved predictive performance of R^2 = 0.985 and 0.643 for reduction and oxidation potentials, respectively. Feature analysis revealed that key descriptors of redox potential were associated with the amplitude of frontier orbital eigenstates, providing valuable guidance for molecular screening. Moreover, to explicitly quantify statistical relationships between molecular structural features and redox potentials, Zhang *et al.*[633] described a GPR model to predict redox potentials based solely on electrolyte additive molecular structures. The model achieved prediction MAEs of 0.05 and 0.10 V for reduction and oxidation potentials, respectively. This model explicitly demonstrated numerical correlations between molecular descriptors and redox potentials.

Organic batteries have recently experienced continuous growth due to their potential advantages, including high capacity, abundant materials, low cost, and environmental sustainability.

The redox properties of organic electrode molecules directly influence the energy density of battery systems. Most organic cathodes exhibit a trade-off between specific capacity and operating voltage, with low redox potentials resulting in insufficient battery voltage for achieving high energy density.[634] To facilitate rapid screening of promising candidates, Allam *et al.*[630] employed feature engineering techniques including LASSO feature selection, relative contribution analysis, and recursive feature elimination to train three learning models: ANN, GBR, and kernel ridge regression (Fig. 10c). These models demonstrated excellent performance in predicting redox potentials for organic molecules outside the original dataset, with electron affinity and number of Li atoms identified as the most critical determinants of redox potential. The methodology provides valuable insights for HTVS processes. Furthermore, Xu *et al.*[631] integrated ML predictions with an autonomous experimental platform for material synthesis and reaction monitoring, successfully developing high-performance organic cathode materials (Fig. 10d). An experimental dataset of 600 entries was compiled, and SMILES representations of molecular structures were converted into MACCS molecular fingerprints to serve as model inputs. A ternary classification framework was implemented using voltage thresholds of 2.5 and 3.5 V, where materials exceeding 3.5 V were deemed optimal. The model achieved an accuracy of 91.6% on the test datasets. SHAP analysis revealed structural preferences for specific functional groups, particularly nitrogen atoms, benzene rings, and carbazole moieties. Guided by computational insights and chemical principles, polymer p-BCz-PH was synthesized through the autonomous platform. The cathode material demonstrated enhanced discharge voltages of 4.5 and 4.8 V *vs.* Li/Li^+, while maintaining high specific capacity throughout charge–discharge cycles.

It is worth noting that frontier molecular orbitals, *i.e.*, the LUMO and HOMO, are commonly employed for high-throughput computational screening due to their easy accessibility by DFT calculations. However, caution is necessary when interpreting LUMO and HOMO energy levels in relation to redox potentials. The former parameters are derived from approximate electronic structure theories when investigating isolated molecular species, and do not explicitly represent the actual participants in redox reactions. In contrast, the redox potential directly correlates with the Gibbs free energy difference between reactants and products in electrochemical processes.[635] While insights into electrolyte stability are provided by both HOMO/LUMO energy levels and redox potentials from different theoretical perspectives, these are constituted as distinct physicochemical descriptors without a strict quantitative correspondence.

4.2. Dielectric constant and donor number

The dielectric constant, originating from physics as a measure of the dielectric properties of materials, has attracted significant attention in solution chemistry due to its critical role in modulating microscopic interactions and solvation structures. The dielectric constant is widely recognized as an important parameter for solvent-mediated regulation of interactions. Specifically, it influences electrostatic interactions among ions, solvents, and dissolved species. For instance, the solubility of salts in solvents is governed by the competition between cation–anion interactions within solid salts and ion–solvent interactions within the solution.

To develop QSPR descriptors correlating with dielectric constants, Cocchi *et al.*[636] established QSPR models for organic solvent dielectric constants using multiple linear regression analysis and multivariate partial least squares (R^2 = 0.956). Schweitzer *et al.*[637] employed ANN modeling on a diverse dataset of 497 compounds with varying dielectric constants, achieving a test set RMSE of 2.33. Subsequent development of specialized models for feature subsets reduced the test error to 1.85.[638] For electrolyte mixture systems, Yao *et al.*[639] predicted dielectric constants at different concentrations using ANN models (Fig. 11a). Ethylene carbonate (EC) content and salt percentage were employed as inputs, while the natural logarithm of dielectric constant was used as the output. After 20 000 training steps, the ANN model demonstrated excellent predictive performance with an MSE of approximately 0.0088. Based on the ANN model, 2D maps illustrating dielectric constant variations with EC percentage and salt concentration in EC/dimethyl carbonate (DMC)/lithiumbis(fluorosulfonyl)imide (LiFSI) electrolytes were generated, showing close agreement with MD simulations. These results demonstrate the remarkable capability of ML models in predicting dielectric constants for complex electrolytes containing multiple solvents and salts.

In polymer-based solid-electrolyte systems, a high dielectric constant is expected to facilitate the dissociation of Li salts, thereby significantly enhancing the ionic conductivity of polymer solid electrolytes.[643] From the viewpoint of molecular representation, polymers can be approximated as assemblies of small molecules. Accordingly, common molecular representation methods have been extended to polymer systems, where strings, molecular descriptors, and molecular fingerprints have been used for characterization.[644–646] Liang *et al.*[640] reported a method for representing polymers at the molecular level and predicting their dielectric constants (Fig. 11b). Polymers were treated as 1D chains, and SMILES was employed to clearly encode specific main-chain and side-chain features, therefore labelling each polymer in the database. A total of 29 features were extracted, encompassing length, quantity, and particular functional groups. RF models were then utilized for ML training, and an autonomous intelligent cloud laboratory was employed to synthesize the predicted polymers.

The limitations of dielectric constant in comprehensively describing Li salt solubility and dissociation characteristics have motivated the development of more holistic descriptors for advanced electrolyte design.[647,648] The DN was initially introduced by Gutmann to characterize the basicity of a solvent and its tendency to donate electrons to an electron acceptor. In Li metal batteries (LMBs), DN has been applied to describe the electron-donating capability of solvents or anions, influencing battery performance primarily through Li salt solubility and the regulation of Li solvation structures. In localized high-concentration electrolyte systems, Chen *et al.*[649] established a DN-based design

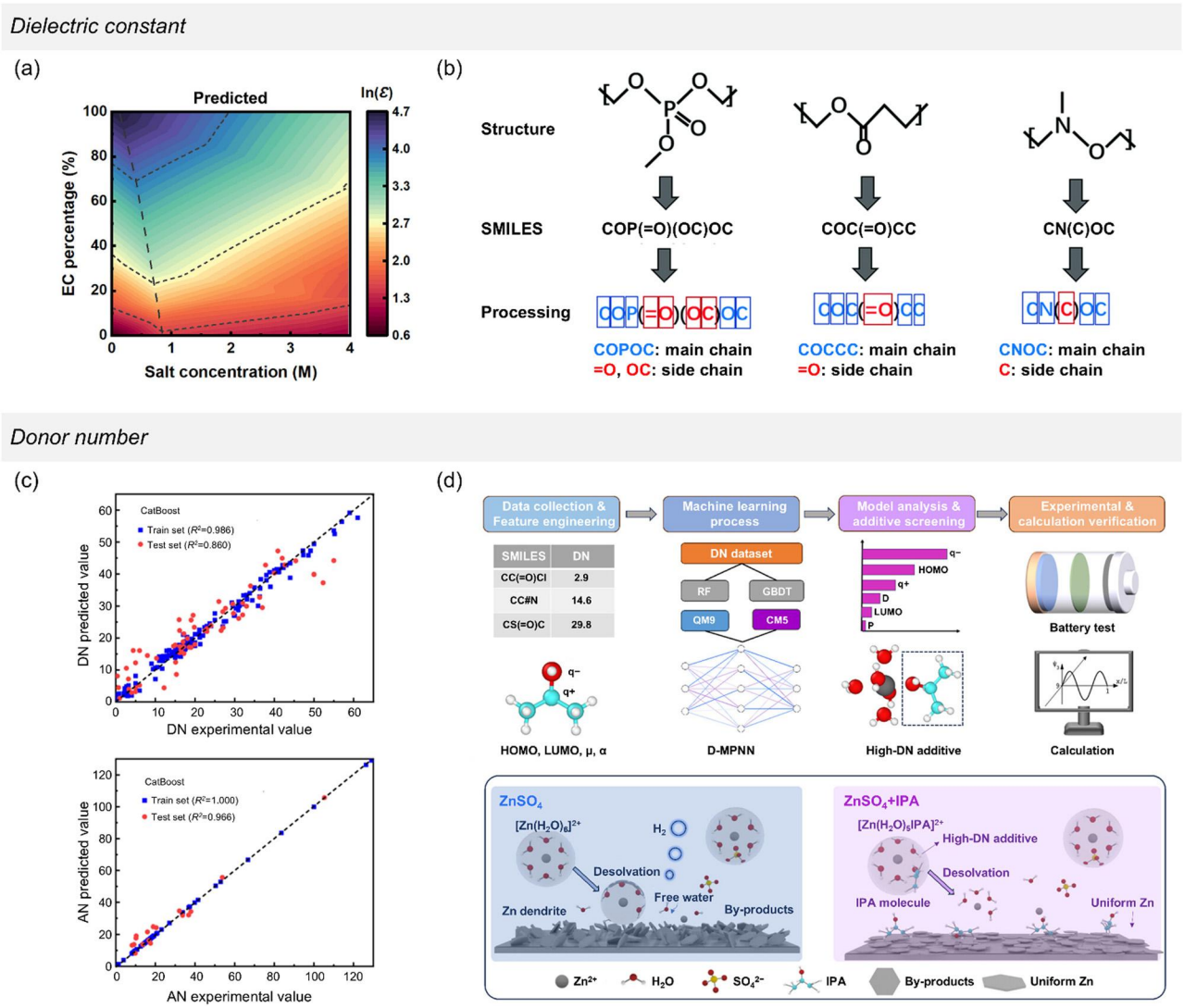

Fig. 11 The prediction of molecular dielectric constants and DNs. (a) ML-predicted dielectric constants of EC/DMC/LiFSI systems under varying EC molar ratios and salt concentrations.[639] Reproduced with permission from ref. 639. Copyright 2021 Wiley-VCH. (b) Polymer electrolyte dielectric constant prediction *via* SMILES-based ML, identifying structural features (main/side chain lengths, functional groups).[640] Reproduced with permission from ref. 640. Copyright 2021 The Royal Society of Chemistry. (c) ML-enabled DN/AN prediction for solvent molecules, with CatBoost model validation against experimental data.[641] Reproduced with permission from ref. 641. Copyright 2024 Elsevier Inc. (d) Data-driven electrolyte additive discovery framework targeting high-donor-number compounds for dendrite-free aqueous Zn-ion batteries.[642] Reproduced with permission from ref. 642. Copyright 2025 American Chemical Society.

criterion requiring primary solvents with DN $>$ 10 paired with diluents possessing DN $\leq$ 10. For S cathode systems, high-DN solvents enhance Li polysulfide solubility and stabilize $S_3^{\bullet-}$ radicals, thereby improving S conversion kinetics.[650–652] Extending this concept to Zn-ion batteries, Cao *et al.*[653] adopted DN as a metric for molecular-cation affinity evaluation, employing high-DN dimethyl sulfoxide solvent additives to suppress Zn dendrite formation and water decomposition at anode/electrolyte interfaces.

To enable accurate DN prediction, Hu *et al.*[641] developed ML models using four algorithms, including CatBoost, GBR, RF, and ridge regression, for simultaneous DN and acceptor number (AN) prediction from molecular descriptors (Fig. 11c). The CatBoost-based models demonstrated superior performance with test set R^2 values of 0.86 (DN) and 0.96 (AN). Addressing Zn-ion battery additive screening, Luo *et al.*[642] created an integrated ML model predicting DN values from molecular fingerprints (Fig. 11d). Experimental validation revealed a direct correlation between higher additive DN values and extended Zn anode calendar life. Isopropanol additives with DN of 36 exhibited strong electrochemical performance in various Zn-based batteries, achieving 1500-hour calendar life, 99% CE over 450 cycles, and superior capacity retention.

To elucidate molecular interactions and solvation structures, nuclear magnetic resonance (NMR) spectroscopy is frequently employed. NMR spectroscopy is a nondestructive, atom-specific technique particularly suited to probing the local chemical environments of nuclei within solvation shells. For

detailed analysis of structure–spectra relationships, Xu *et al.*[654] introduced the NMRNet framework, in which a novel SE(3)-equivariant Transformer architecture was employed to predict liquid and solid-state NMR chemical shifts with impressive performance. NMRNet supports both single-nucleus and multi-nucleus prediction and has demonstrated superior performance across multiple evaluation metrics, thereby providing a powerful tool for molecular structure elucidation and molecular design. Then, You *et al.*[655] developed a ML-augmented method for dynamic prediction of ^{7}Li chemical shifts in LiFSI/1,2-dimethoxyethane electrolytes. A reversal in ^{7}Li chemical shift trends was observed;[656] an upfield shift occurred as LiFSI concentration increased from 1 M to 3 M, followed by a downfield shift at 4 M, which is consistent with the experimental results. The quantitative mapping between molecular structure and NMR spectroscopy has paved the way for optimized electrolyte design.

4.3. Viscosity and ionic conductivity

Viscosity is defined as the resistance of a fluid to flow under applied shear, which originates from internal friction arising between adjacent fluid layers moving at different relative velocities. In working electrolytes, viscosity represents a crucial characteristic, directly influencing ionic transport behavior. Electrolytes with optimized viscosity are particularly essential for applications operating under harsh conditions such as low temperatures or fast-charging scenarios. Macroscopically, viscosity affects the wettability of the electrolyte onto the separator and cathode in assembled batteries. High-viscosity liquids generally exhibit slower wetting or spreading on solid surfaces compared to low-viscosity fluids. Therefore, regulating viscosity is vital for practical electrolyte deployment.[657]

Viscosity prediction methods have progressed from single-component systems at room temperature to multi-component systems across variable temperatures. For instance, Goussard *et al.*[658] proposed a ML model capable of predicting viscosities at 25 °C for 300 pure organic liquid compounds. Extending temperature applicability, Chew *et al.*[659] introduced descriptor-based and GNN models to predict viscosities across different temperatures (Fig. 12a). A comprehensive dataset consisting of viscosities for over 4000 organic small molecules was established. MD-derived descriptors can capture intermolecular interactions and improve viscosity prediction accuracy, particularly under data-limited conditions. The ML models successfully captured the inverse relationship between viscosity and temperature for six potential co-solvents suitable for LIBs, including methyl acetate and ethyl acetate. Furthermore, Bilodeau *et al.*[660] proposed an automated pipeline to predict the viscosity of liquid mixtures while accounting for temperature effects (Fig. 12b). A substantial dataset comprising 1734 compounds and 39 077 mixture data points from the literature was assembled. The model, based on a directed MPNN architecture using molecular graphs, mole fractions, and temperature as inputs, achieved a MAE of 0.043 in log(cP) units.

Ionic liquids (ILs) have emerged as promising electrolyte solvents due to their enhanced stability and non-flammability compared to conventional organic solvents.[664] Typically, the viscosity of ILs is one to two orders of magnitude higher than that of routine organic liquids. The high viscosity and consequently moderate ionic conductivity severely limit the battery cycling performance of IL-based electrolytes at room temperature.[665] Numerous studies focused on predicting the viscosity of ILs. For instance, Han *et al.* utilized multiple linear regression to correlate the viscosities of imidazolium-based ILs at 298.15 K. Zhao *et al.*[666] constructed two predictive models, a multi-linear regression model and an SVM model, using 1079 experimental viscosity data points for 45 imidazolium-based ILs, measured under pressures ranging from 1 to 3000 bar and temperatures between 273.15 and 395.32 K. The SVM model achieved an impressive R^2 value of 0.977. Huwaimel *et al.*[667] collected a comprehensive dataset comprising 8500 entries to develop predictive models for the viscosity of IL-containing mixtures. Employing input variables such as cation and anion types, temperature, and IL concentration, the RF model attained an exceptionally high R^2 score of 0.997.

Ionic conductivity, as a fundamental parameter characterizing the efficiency of ionic transport within electrolytes, reflects the directional migration capability of ions under applied electric fields. Its value is collectively determined by carrier concentration, ionic charge, and mobility, directly influencing electrolyte performance.[668] Shi *et al.*[661] built linear and nonlinear QSPR models based on molecular descriptors to predict the ionic conductivity of ILs, explicitly accounting for temperature effects (Fig. 12c). As a result, employing the ion-pair representation with a back-propagation ANN yielded a test-set R^2 of 0.989 without any notable outliers, consistent with the findings reported by Yang *et al.*[669] Additionally, Chen *et al.*[662] combined physically derived COSMO-RS modeling with robust ML techniques, effectively reducing the MAE of purely physical models from 0.550 to 0.396 (Fig. 12d). The dependencies of ionic conductivity on temperature, alkyl chain length, cation alkyl-chain branching, and anion volume were emphasized, providing valuable insights for selecting and developing predictive models based on these IL properties and guiding other molecular property predictions.

Furthermore, the relatively low ionic conductivity of solid polymer electrolytes (SPEs) compared with liquid and ceramic solid electrolytes has limited their widespread adoption in functional battery systems. To accelerate SPE molecular discovery, Bradford *et al.*[663] introduced ChemArr, a neural network trained on data from 217 experimental publications, to predict SPE ionic conductivity (Fig. 12e). ChemArr employs Chemprop's directed MPNN to encode polymer and salt molecular graphs, integrates salt concentration and polymer molecular weight, and then outputs Arrhenius parameters for conductivity prediction. It was further applied to over 20 000 hypothetical SPE formulations derived from 820 synthetic polymers. A 2D UMAP projection of the predicted polymer space revealed clusters of polymers with high predicted conductivity, providing a valuable roadmap for guiding future experimental efforts (Fig. 12f).

4.4. Melting, boiling, and flash points

The exploration of safe batteries operated under extreme conditions (*e.g.*, Arctic, desert conditions) critically depends on

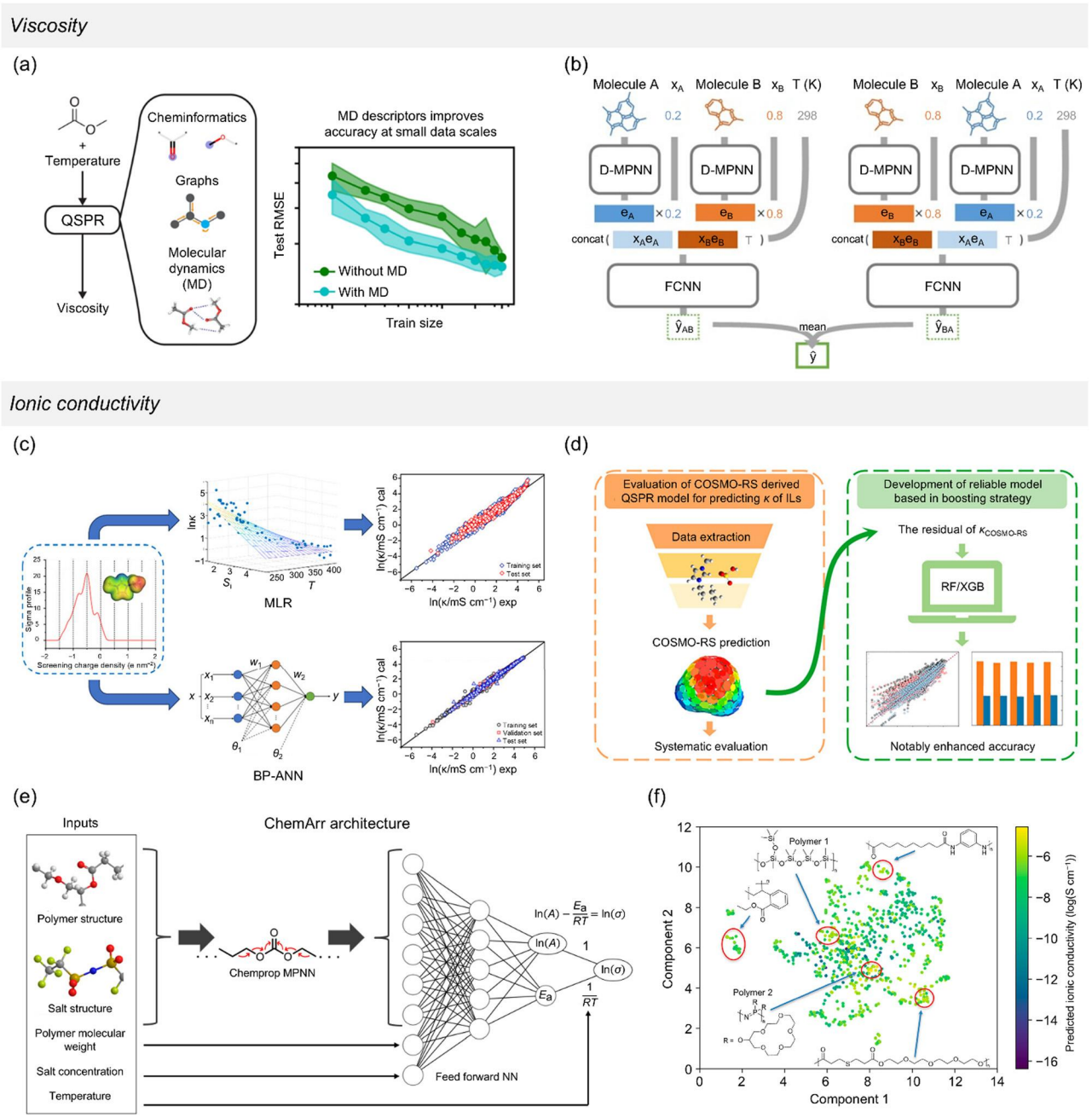

Fig. 12 The prediction of molecular viscosity and ionic conductivity. (a) Physics-informed ML models enhance viscosity prediction accuracy.[659] Reproduced with permission from ref. 659. Copyright 2024 The Authors. (b) Binary mixture viscosity prediction employs directed MPNN modified for dual-component systems where molecule A/B embeddings (e_A/e_B) process structural features with mole fractions x_A/x_B.[660] Reproduced with permission from ref. 660. Copyright 2023 Elsevier Inc. (c) QSPR models combine multiple linear regression and ANNs for IL conductivity prediction.[661] Reproduced with permission from ref. 661. Copyright 2024 American Chemical Society. (d) Conductivity prediction integrates COSMO-RS derived QSPR with boosted ML algorithms.[662] Reproduced with permission from ref. 662. Copyright 2024 American Chemical Society. (e) Chemistry-informed ML (ChemArr) architecture predicts polymer electrolyte conductivity by concatenating MPNN-generated molecular features with numerical parameters to solve Arrhenius equations.[663] (f) ChemArr-based high-conductivity polymer screening visualizes the design space where colored data points represent predicted ionic conductivities of polymers at 25 °C, highlighting optimal structural regions with annotated examples.[663] Reproduced with permission from ref. 663. Copyright 2023 American Chemical Society.

precise prediction of melting,[670–679] boiling,[680–685] and flash points,[686–688] necessitating comprehensive high-quality thermophysical datasets to enable reliable computational modeling frameworks. Bergström *et al.*[689] constructed a dataset containing 277 molecules for melting point prediction and applied partial least squares projection methods, achieving an RMSE of 44.6 K. The dataset has since been widely used as a benchmark for comparative melting point prediction studies. Subsequently,

Karthikeyan *et al.*[690] expanded the dataset significantly, compiling a diversified set of 4173 compounds. A FNN model was established and compared with Bergström's original dataset, achieving an MAE of 32.6 K. Tetko *et al.*[691] further substantially extended the available melting point data, compiling published datasets comprising melting points for over 47 000 compounds and subsequently extracting nearly 300 000 data points from patent literature, greatly enriching the melting point dataset.[692]

For boiling point prediction, Needham *et al.*[693] created a dataset of 74 alkanes, reporting R^2 = 0.999. Balaban *et al.*[694] expanded this approach to 532 halogenated hydrocarbons, employing molecular descriptors and achieving R^2 = 0.97. Katritzky *et al.*[695] constructed an initial dataset of 298 diverse organic compounds to predict normal boiling points (R^2 = 0.973), subsequently enlarging it to 612 compounds (R^2 = 0.965).[696] The data size was further expanded by Gharagheizi *et al.*[697] FNNs were utilized to predict the normal boiling points of a considerably larger dataset containing 17 768 compounds, and an RMSE of 21 K was achieved on the test set. More recently, Qu *et al.*[698] leveraged a dataset of 22 935 experimental data points to develop GCN models, attaining an MAE below 6 K.

In flash point prediction, Katritzky *et al.*[699] applied ML models on 271 diverse compounds, achieving an R^2 value of 0.953. Importantly, a strong correlation between flash points and experimental or predicted boiling points was demonstrated, providing foundational guidance for subsequent research utilizing boiling points as descriptors for flash point prediction.[700–702] Then, the dataset was extended to 758 organic compounds, and ANN models incorporating geometric, topological, quantum-mechanical, and electronic descriptors were developed, achieving an R^2 of 0.978 and an MAE of 12.6 K.[703] Additionally, Zhokhova *et al.*[704] developed a database of flash points for 525 organic compounds, employing linear models based on fragment descriptors and ANN models, resulting in an R^2 of 0.959 and an RMSE of 14.6 K. Gharagheizi *et al.*[705] further predicted flash points of 1378 organic compounds using functional group-based neural network models, achieving an excellent R^2 of 0.97 and a standard error of prediction of 13.1 K. Subsequently, Le *et al.*[706] assembled an extensive dataset comprising 9399 chemically diverse compounds, with flash points ranging from below −130 to above 900 °C. Bayesian regularized ANNs with a Laplacian prior were utilized, and an R^2 value of 0.95 was achieved.

However, most current reported methods remain confined to predicting a single property. For simultaneous prediction of electrolyte properties, Gao *et al.*[86] proposed a knowledge-based electrolyte property prediction integration (KPI) achieving MAEs of 10.4, 4.6, and 4.8 K for melting, boiling, and flash points, respectively (Fig. 13). The framework outperformed existing models in 18/20 benchmark datasets by systematically correlating molecular structures with thermophysical properties. Leveraging chemical domain knowledge, KPI combined molecular neighborhood analysis and HTVS to identify 29 potential electrolyte candidates suitable for extreme-temperature battery applications. The knowledge–data dual-driven methodology demonstrates superior predictive accuracy while elucidating fundamental structure–property relationships, particularly through its integration of explainable AI with chemical intuition.

5. Applications to molecular design in rechargeable batteries

While significant progress has been achieved in property prediction, the transition from understanding to electrolyte design remains a central challenge, due to the vastness of chemical space and the complexity of structure–property relationships. With the advancement of AI, the paradigm of molecular design has shifted from empirical trial-and-error toward data-driven and knowledge-guided strategies.

In this section, the applications of AI technologies to molecular design are discussed from multiple perspectives. Interpretable models are first introduced to enable knowledge discovery and guide rational design. Then, HTVS methods are presented for efficiently navigating chemical space. Oriented molecular generation is highlighted as a tool for directly creating structures with target properties. Finally, the integration of active learning and autonomous laboratories is illustrated, demonstrating how closed-loop experimentation accelerates the discovery and optimization of high-performance molecules for rechargeable batteries.

5.1. Knowledge discovery

5.1.1. Concepts of interpretable machine learning. One of the fundamental objectives of chemical research is to establish generalizable structure–property relationships and mechanistic models, essentially abstracting and generalizing hidden patterns within complex systems. Traditional ML, particularly DL, has demonstrated remarkable predictive capabilities. However, the inherent black-box nature of these models often creates a disconnect between model predictions and the underlying chemical mechanism. The conflict has motivated researchers to reconsider the essential value of ML models, not merely as predictive tools, but also as vehicles for knowledge discovery. Consequently, IML has emerged as a leading frontier in the interdisciplinary fields of cheminformatics and computational materials science. The central objective of IML is to construct transparent or traceable models that transform data-driven statistical correlations into physically and chemically meaningful, interpretable knowledge.

IML can be broadly divided into two approaches: intrinsically interpretable models and *post hoc* interpretability methods. The former emphasizes simplicity and transparency, designing models (*e.g.*, linear regression, decision trees, rule lists) whose parameters directly map onto input features. Linear regression clearly associates regression coefficients with physical variables, while decision trees and rule lists produce easily interpretable rules reflecting specific molecular or structural boundaries. However, these models often assume linearity or piecewise linearity, limiting their ability to capture the nonlinear interactions commonly observed in battery systems.

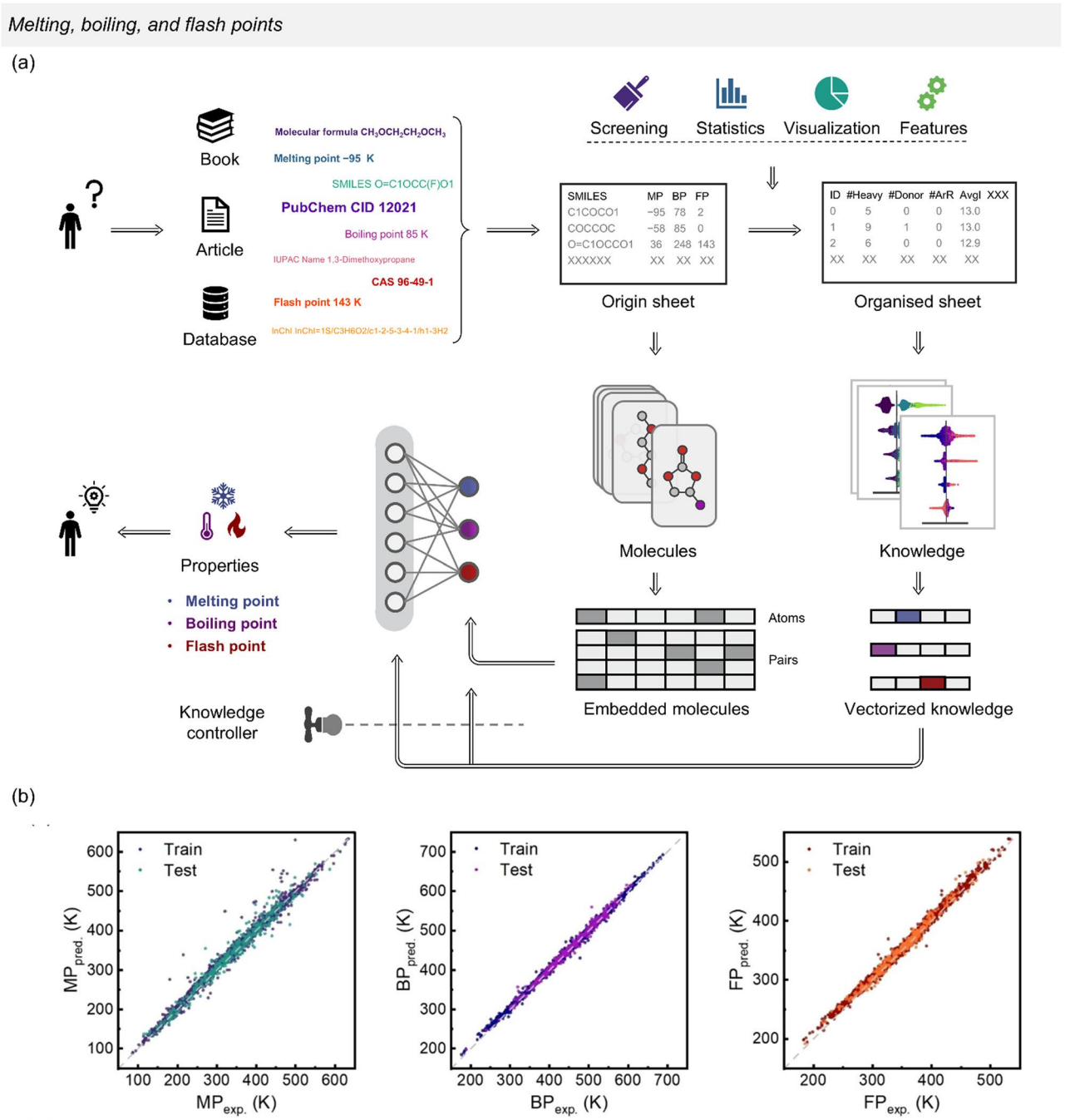

Fig. 13 The knowledge–data dual-driven electrolyte property prediction. (a) Framework integrating domain knowledge and data predicts electrolyte properties (melting, boiling, and flash points).[86] (b) Prediction results for the state-of-the-art models of melting points (left), boiling points (medium), and flash points (right).[86] Reproduced with permission from ref. 86. Copyright 2024 Wiley-VCH.

Post hoc interpretability techniques aim to mitigate these limitations by providing human-understandable explanations for complex, often black-box models. Feature attribution analyses, including SHAP and local interpretable model-agnostic explanations, quantify each feature's contribution to the prediction, either globally (*e.g.*, *via* Shapley values) or locally (*e.g.*, by constructing linear approximations around individual samples). Surrogate models and diagnostics, such as partial dependence plots, reveal marginal effects of specific features by perturbing the remaining variables, thereby highlighting potential nonlinear behaviors. Furthermore, visualization mappings offer additional insights. By projecting high-dimensional features or network states into lower-dimensional visual spaces, these methods pinpoint the most influential input regions or interactions.

5.1.2. Knowledge discovery guiding molecular rational design. The strategic application of intrinsically interpretable models has enabled fundamental discoveries in electrolyte

optimization. Kim *et al.*[88] utilized highly interpretable models to uncover the relationship between electrolyte molecular structures and CE (Fig. 14a). Experimental data from over 150 Li||Cu battery tests were collected, encompassing electrolyte systems including conventional, high-salt, localized high-concentration, fluorinated, dual-salt, and additive-enhanced formulations, with a CE range from 80% to 99.5%. By constructing molecular descriptors for the collected data and employing linear models, negative correlations of the oxygen content ratio in solvents and the carbon content ratio in anions with CE were observed. In contrast, high proportions of inorganic components and fluorine-to-oxygen ratios positively influenced CE. Notably, contrary to the traditionally emphasized role of fluorinated molecules, the oxygen of solvents was the most decisive factor affecting CE (Fig. 14b). While fluorination of solvent molecules is known to weaken Li^+ solvation, it was inferred that similar effects were achieved by reducing oxygen content in solvents. Guided by these fresh insights, a fluorine-free solvent electrolyte was proposed, rendering a high CE of 99.7%.

Complementing these analytical approaches, visualization-driven strategies are gaining traction in molecular discovery. Li *et al.*[707] adopted a dimension-reduction-based clustering visualization strategy to screen additives for aqueous Zn-ion battery electrolytes (Fig. 14c). Specifically, a random tree embedding algorithm was utilized to transform 17 physical descriptors, after which *k*-means clustering algorithm and Voronoi partitioning

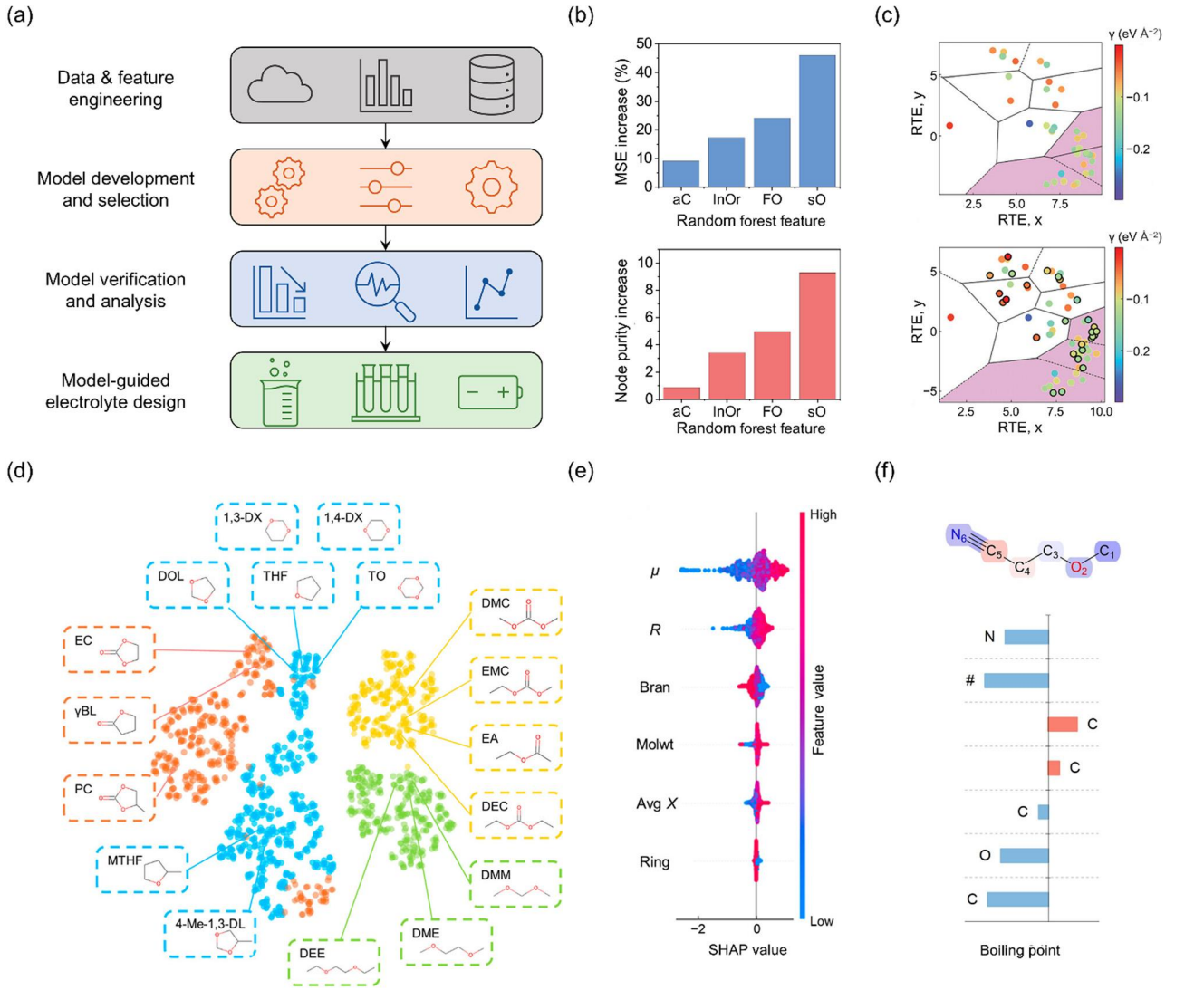

Fig. 14 Knowledge discovery guiding molecular rational design. (a) Li metal anode electrolyte design *via* automated data acquisition and model development.[88] (b) Feature importance analysis using MSE and node purity metrics.[88] Reproduced with permission from ref. 88. Copyright 2023 The Authors. (c) AI-driven additive selection for aqueous Zn-ion batteries employing RTE dimensionality reduction and Voronoi clustering.[707] Reproduced with permission from ref. 707. Copyright 2024 Wiley-VCH. (d) Solvent database generation and visualization *via* clustering methods to reveal the reductive stability of ion–solvent complexes in Li battery electrolytes.[85] (e) IML prediction of ion–solvent complex LUMO energy levels with SHAP-based ether feature ranking.[85] Reproduced with permission from ref. 85. Copyright 2023 American Chemical Society. (f) Intelligent screening of wide-temperature electrolytes through atom-level explainable analysis of boiling points.[708] Reproduced with permission from ref. 708. Copyright 2024 Wiley-VCH.

were applied in the 2D space. Using silhouette analysis, the optimal cluster number was determined to be nine, enabling the identification of regions most significantly associated with enhanced battery stability. Within these identified regions, molecules such as 1,2,3-butanetriol and acetone exhibited notably low predicted surface free energies, indicative of structural stability, thereby emerging as promising additive candidates. Subsequent experiments demonstrated that batteries incorporating acetone and 1,2,3-butanetriol as electrolyte additives significantly outperformed pure $ZnSO_4$ electrolytes in terms of initial CE, voltage polarization, and cycling stability.

There are increasing contributions that employed *post hoc* SHAP analysis to decode complex molecular interactions. To explore factors influencing the reduction stability of electrolyte solvent molecules, Gao *et al.*[85] proposed a graph-theory-based molecular generation method, constructing a database comprising 1399 electrolyte solvent molecules (Fig. 14d). First-principles calculations were employed to determine the LUMO energy levels of these electrolyte molecules. The study revealed that the reduction stability of electrolytes decreased when solvent molecules formed ion–solvent structures. A combination of RF modeling and Shapley value-based interpretability analysis was utilized, and molecular dipole moment and molecular radius were identified as critical descriptors influencing electrolyte reduction stability (Fig. 14e). The data-driven approach not only investigated the reduction stability of electrolyte ion–solvent structures but also uncovered essential factors governing electrolyte stability, thus providing valuable theoretical insights into advanced electrolyte molecule design.

Beyond feature attribution methods, surrogate modeling approaches have proven effective in bridging DL predictions with chemical intuition. Qin *et al.*[708] employed surrogate models to extract interpretable knowledge embedded within DL black-box models, facilitating the design of non-aqueous electrolytes suitable for wide temperature operation. Interpretability analysis revealed similarities between nitrile groups (–CN) and fluorine substituents in influencing electrolyte properties, such as boiling point and dielectric constant (Fig. 14f). Taking 3-methoxypropionitrile (MPN) as an illustrative example, the contribution of each atomic site within MPN toward boiling and melting points was quantitatively assessed using interpretable methods. The strongly polar nitrile group positively contributed to elevated boiling and melting points, whereas the presence of ether linkages (−COC−) exhibited an opposing effect, effectively offsetting the increase in melting point and broadening the liquid-phase operating range. With additional introduction of ether linkages, MPN was eventually identified as the primary electrolyte solvent. The MPN electrolyte enabled LIBs to operate reliably over an exceptionally wide temperature range from −60 to 120 °C. Notably, $LiCoO_2$||Li cells utilizing the proposed wide-temperature MPN electrolyte demonstrated stable cycling performance, maintaining a high capacity retention of 72.3% after 50 cycles at 100 °C.

5.2. High-throughput virtual screening

5.2.1. Concepts of chemical space and high-throughput virtual screening. In molecular design for rechargeable batteries, a core challenge arises from the enormous scale of chemical space contrasted with the limited throughput of experimental validation. Chemical space, encompassing all possible molecules and their corresponding properties, expands combinatorially with dimensions determined by atomic types, bonding patterns, and configurational arrangements. The number of theoretically possible organic molecules reaches an astronomical scale of approximately 10^{60}, far exceeding the practical exploration capabilities of experimental techniques.[709] Traditional trial-and-error methods exhibit extremely low search efficiency in such high-dimensional spaces. In contrast, HTVS, employing computationally driven hierarchical optimization strategies, transforms random searches into directed molecular design, and is expected to accelerate the process of molecular discovery.

The workflow of HTVS begins with the rational construction of chemical spaces. Based on functional requirements of target battery systems, molecular space boundaries are defined through combinatorial rules or generative models, constraining molecular diversity to maintain computational feasibility. A multi-step screening process follows. The initial coarse-screening phase prioritizes low computational cost descriptors or cheminformatics toolkits, rapidly filtering molecules, such as eliminating structurally unstable candidates based on topological rules or assessing thermodynamic stability using semi-empirical quantum chemical methods, to substantially reduce the molecular space size. The preliminary screening efficiently excludes unsuitable candidates, typically compressing the initial molecular space to a manageable scale with minimal resource consumption. The subsequent precise-screening stage employs high-accuracy computational methods for detailed validation of targeted properties. QM calculations, particularly those based on DFT, reveal electronic structural characteristics, while MD simulations quantify thermodynamic and transport parameters. Integrating these approaches enables the identification of candidate molecules exhibiting both thermodynamic stability and desired functionalities. Concurrently, AI-driven predictive models significantly accelerate the screening process by establishing mappings between molecular structures and macroscopic properties. Data-driven predictive models rapidly pinpoint potential high-performance candidates, with high-precision computations further validating their reliability.

5.2.2. High-throughput virtual screening accelerating the discovery of advanced molecules. HTVS can significantly accelerate the discovery of high-performance recipes through systematic integration of computational modeling, data mining, and experimental validation. For example, Cheng *et al.*[90] established a foundational hierarchical computational screening framework, conducting multi-stage property evaluations *via* QM calculations for over 1400 organic molecules intended for non-aqueous flow batteries (Fig. 15a). During the initial screening (redox potential evaluation), the number of candidate molecules was reduced from 1417 to 353. The second step (solubility assessment) further decreased the number from 353 to 262. Subsequently, structural stability screening eliminated additional molecules, primarily thiane derivatives susceptible to ring-opening reactions during reduction, ultimately yielding 231 promising candidates. These remaining molecules,

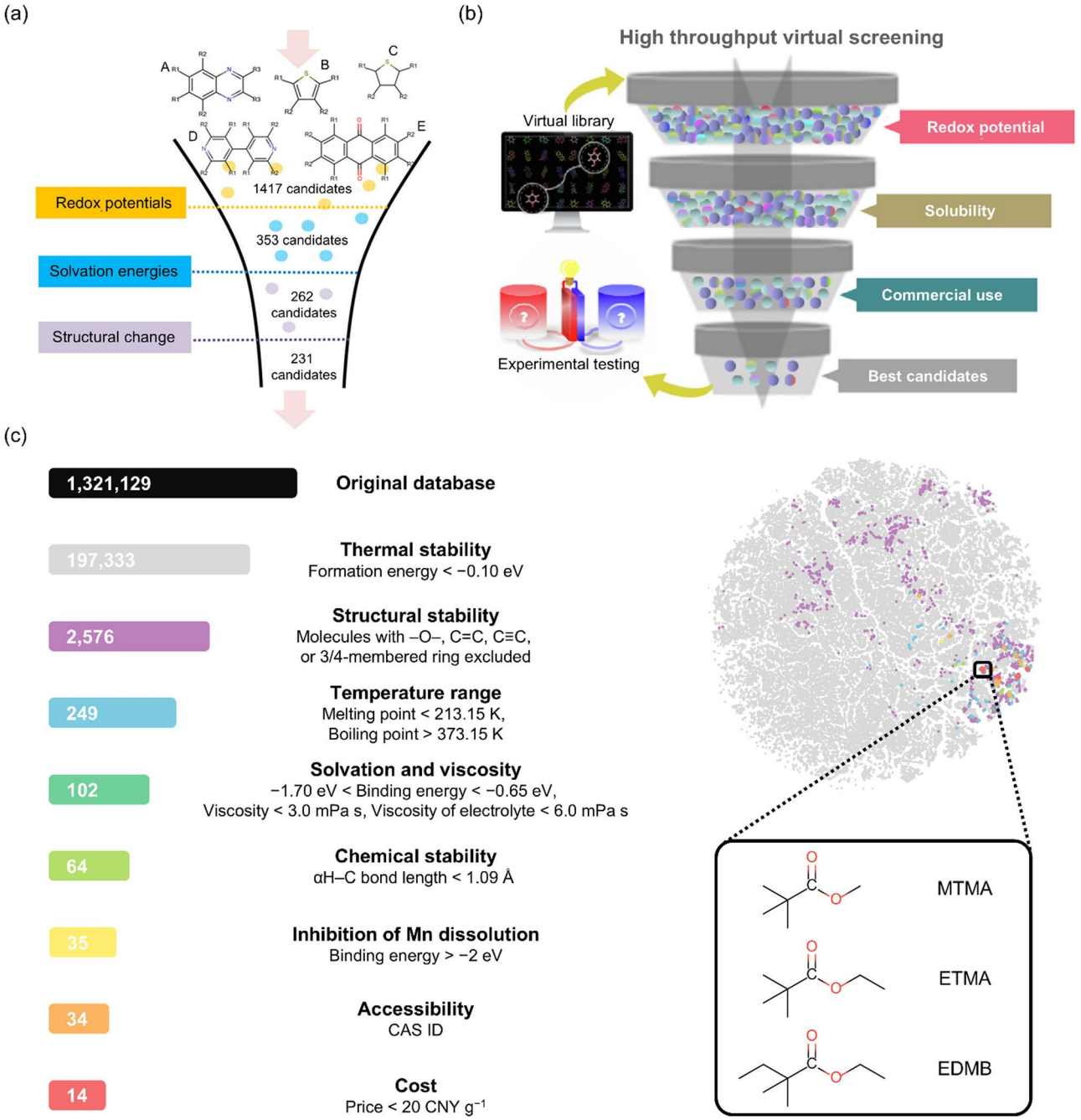

Fig. 15 HTVS accelerating the discovery of advanced molecules. (a) HTVS enables rapid electrolyte discovery for energy storage systems.[90] Reproduced with permission from ref. 90. Copyright 2015 American Chemical Society. (b) Data-driven identification of small electroactive molecules optimizes aqueous redox flow battery performance.[710] Reproduced with permission from ref. 710. Copyright 2022 The Authors. (c) The design of molecules for high-temperature, fast-charging LIB electrolytes was guided by a screening procedure based on predefined physicochemical property requirements for both molecules and resulting electrolytes. Candidate selection was conducted through systematic evaluation and visualization of key molecular and electrolyte performance metrics.[711] Reproduced with permission from ref. 711. Copyright 2025 Wiley-VCH.

satisfying redox window criteria, emerged as suitable candidates for experimental validation. Further advancing the field, Zhang *et al.*[710] employed a data-driven strategy to expand the scope of HTVS (Fig. 15b). A virtual library containing 3257 quinone derivatives for aqueous flow batteries was developed. DFT calculations predicted redox potentials, supervised ML models estimated aqueous solubility, and automated searches within the ZINC database identified commercially available compounds. From the approach, 205 candidate compounds exhibiting superior predicted solubility and lower redox potentials were selected. Among the 205 candidates, 16 commercially available compounds underwent experimental evaluation, with electrochemical characterization performed on seven molecules. As a result, Indigo-3(SO_3H) exhibited notably enhanced solubility,

capacity retention, and CE relative to the benchmark anthraquinone-2,7-disulfonic acid.

For the molecular design of electrolytes targeting fast-charging LIBs, Yang *et al.*[711] proposed a data–knowledge-dual-driven approach, integrating high-throughput calculations, ML techniques, and experimental validation (Fig. 15c). A molecular dataset comprising 1 321 129 compounds was constructed, from which 54 202 candidates were systematically screened using DFT calculations and MD simulations. Key molecular descriptors, encompassing thermal and structural stability, temperature range, solvation, viscosity, and chemical stability, were employed in conjunction with ML models for efficient property prediction. Through this screening process, three novel carboxylate solvents, methyl trimethylacetate, ethyl trimethylacetate (ETMA), and ethyl 2,2-dimethylbutanoate, were identified. Notably, ETMA-based electrolytes demonstrated excellent fast-charging performance, achieving the highest voltage (4.3 V), the highest charging rate (4.0 C), and the longest cycle life (over 4100 cycles) compared with literature reports. For LMBs, Jia *et al.*[456] developed an HTVS strategy for fluorinated ether electrolytes. A database of 5576 candidates was generated from 1510 solvents and four salts. A voting ensemble model using five key descriptors enabled rapid property prediction. A GCN further accelerated descriptor estimation. The optimal molecule achieved stable cycling between 2.8 and 4.4 V *vs.* Li/Li^+ at C/2 with 99.5% CE maintained over 100 cycles.

Beyond electrolyte molecules, Du *et al.*[712] successfully extended HTVS strategies to electrode molecules, specifically targeting carbonyl-based organic electrode molecules. Initially, one million organic molecules were collected from PubChem. Subsequently, considering atomic types, active-site counts, and hierarchical clustering based on previously reported organic electrode molecule characteristics, 1524 molecules were selected as potential candidates. High-throughput calculations determined reduction potentials for a randomly chosen subset of 1200 molecules, forming the training set for an SVR model. Through the methodology, naphthalene-1,4,5,8-tetraone emerged as a molecule with high reduction potential and energy density. Experimental validation confirmed its superior performance, achieving prolonged cycling stability of 2500 cycles at 1 A g^{-1} and a high discharge voltage of 2.5 V. The methodology provides valuable insights for accurately and rapidly screening advanced organic electrode molecules for LIBs.

5.3. Oriented molecular generation

5.3.1. Concepts of generative models and oriented molecular generation. Beyond HTVS, molecules can be custom-designed to meet specific performance targets by leveraging goal-oriented generative models. In this paradigm, models learn latent representations of molecular structures and integrate property predictors, creating a closed loop of iterative generation, evaluation, and optimization. Generative adversarial networks (GANs), variational autoencoders (VAEs), and diffusion models constitute key approaches.

GANs employ generator and discriminator networks within a game-theoretic framework. The generator produces candidate molecules, while the discriminator distinguishes real from synthesized data. Conditional GANs introduce property labels, guiding the generator toward desired chemical or electrochemical characteristics. However, challenges such as mode collapse and discrete-space backpropagation remain. VAEs rely on an encoder–decoder architecture that encodes molecules into continuous latent spaces and then reconstructs them. Property tuning is enabled through latent interpolation or optimization due to the continuity. Conditional VAEs further integrate property constraints, directing the decoder to generate molecules with specific attributes, such as high ionic conductivity. Nonetheless, latent space regularization can occasionally produce conservative designs. Diffusion models follow a Markov-chain approach, progressively adding noise to data and learning a reverse denoising process to reconstruct molecules. Stable training is offered, and structurally complex, multi-ring molecules are effectively generated. Moreover, SE(3)-equivariant diffusion models precisely handle 3D configurations. Through conditional guidance, these models incorporate property requirements, yielding molecules aligned with predefined design objectives. By combining generative models with predictive property constraints, one can transcend empirical exploration and actively engineer molecules tailored to specialized tasks, including electrolyte additives and active electrode components. The approach accelerates innovation in advanced battery materials and related systems by systematically navigating the vast chemical space and directing molecular evolution toward optimal performance.

5.3.2. Oriented molecular generation promoting precise molecular design. In the oriented molecular design of liquid organic molecules, Tagade *et al.*[713] proposed a DL inverse prediction framework named SLAMDUNCS, employing novel conditional sampling strategies for property-driven molecular design. A binary representation was developed to digitally encode molecular structures, and semi-supervised learning methods were applied to establish structure–property mappings. SLAMDUNCS predicted molecules exhibiting reduction potentials lower than −3.35 V relative to the standard hydrogen electrode. Validation of randomly selected 50 candidate molecules demonstrated a maximum prediction error of 0.59 V and an MAE of 0.20 V, compared with DFT calculations.

In designing SPEs, Yang *et al.*[714] leveraged generative AI for *de novo* polymer discovery, successfully identifying candidates with superior ionic conductivity (Fig. 16a). In unconditional generation tasks, minGPT and diffusion-LM models exhibited excellent performance in producing novel, valid, and synthesizable polymers, while the 1D diffusion model showed comparatively limited efficacy. The minGPT demonstrated superior replication of polymer property distributions and proved computationally more efficient. In conditional generation experiments targeting ionic conductivity, minGPT-generated polymers exhibited conductivity distributions distinctly skewed toward high values. MD simulations of 50 selected polymers (from 100 000 high-conductivity-labeled candidates) resulted in successful conductivity validation for 46 candidates, with 17 exhibiting conductivity values exceeding all polymers in the original training set, some even doubling the benchmark values. Khajeh *et al.*[715]

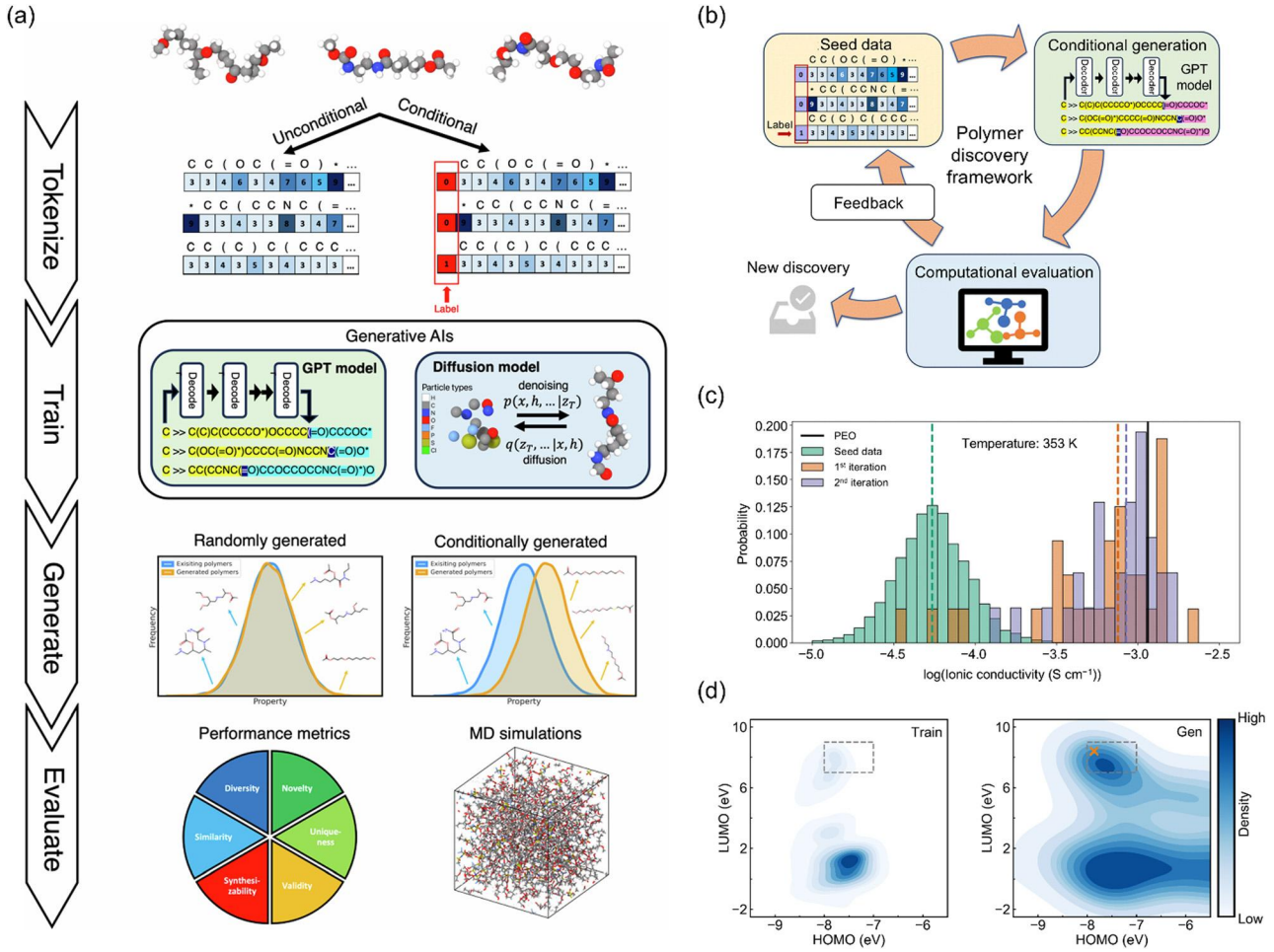

Fig. 16 Oriented molecular generation promoting precise molecular design. (a) *De novo* polymer electrolyte design integrates GPT-based and diffusion models, processing SMILES representations from HTP-MD datasets through label-free and property-guided generation. High-conductivity candidates undergo MD validation.[714] Reproduced with permission from ref. 714. Copyright 2024 The Authors. (b) A conditional generative framework optimizes polymer electrolyte discovery.[715] (c) MD-predicted ionic conductivity distributions show iterative improvements (training set: green; first iteration: orange; second iteration: purple), with dashed lines indicating distribution means and the black line marking the polyethylene oxide.[715] Reproduced with permission from ref. 715. Copyright 2024 The Authors. (d) The Uni-Electrolyte AI platform demonstrates out-of-domain HOMO–LUMO prediction capability, generating dense molecular distributions in target regions (including DMC absent in training data) for rechargeable battery electrolyte design.[716] Reproduced with permission from ref. 716. Copyright 2025 Wiley-VCH.

further proposed a polymer discovery framework comprising three core components: a conditional generative model (minGPT-based), a computational evaluation module (MD simulations), and a feedback mechanism (Fig. 16b). Through iterative training and strategic sampling, the framework identified 14 polymers surpassing polyethylene oxide in ionic conductivity (Fig. 16c).

The establishment of an electrolyte molecule development platform integrated with molecular generation tools holds great promise for accelerating the molecular discovery process and making it accessible to a broader community of researchers. Chen *et al.*[716] introduced an AI-based platform, Uni-Electrolyte, dedicated to the design of electrolyte molecules for rechargeable batteries, consisting of three interconnected modules: EMolCurator, EMolForger, and EMolNetKnittor. EMolCurator constructs databases using DFT calculations and MD simulations for molecular property predictions, supporting multi-criteria screening, similarity searches, and AI-driven molecule generation. EMolForger employs GNNs and reaction planners to predict synthesis routes and optimize reaction conditions. EMolNetKnittor leverages stochastic kinetic Monte Carlo simulations and proprietary databases to analyze SEI formation and predict decomposition products. For instance, the conditional diffusion model accurately generated electrolyte molecules with specified HOMO–LUMO gaps. In cases such as dimethoxyethane, despite sparse representation in the initial dataset, the conditional diffusion model successfully generated molecules closely aligned with the targeted HOMO–LUMO gap region, demonstrating its sensitivity to conditional guidance and capability to explore sparsely populated chemical spaces (Fig. 16d). These achievements highlight the significant practical utility and potential of AI-driven generative models for designing electrolyte molecules

with tailored properties, thus greatly facilitating the development of advanced battery materials.

5.4. High-throughput experimentation

5.4.1. Intelligent robotic system and autonomous laboratory. Autonomous chemistry laboratories represent a novel paradigm that integrates AI and automation, rooted in the evolution of HTE. Advances in ML have ushered in data-driven decision-making, transforming automated experimentation into a genuinely intelligent process. In 2020, the first mobile robotic chemist was constructed at the University of Liverpool under the direction of Andrew Cooper.[717] Over eight days, 688 experiments were autonomously executed, culminating in the discovery of a novel catalytic material. In 2021, the first all-round AI-Chemist with a scientific mind was established at the University of Science and Technology of China by Jun Jiang's team.[718] AI-Chemist was composed of a service platform, a mobile robot, multiple workstations, and a computational brain. Furthermore, within five weeks, over 3.76 million formulations were screened, and a practical oxygen-evolution electrocatalyst was synthesized using Martian meteorite feedstock. All procedures were executed automatically without human intervention.[719] In addition, lots of intelligent robotic systems and autonomous laboratories have been established, significantly accelerating the pace of material innovation.[620,621,720–722]

The autonomous laboratory arises from systematic innovation in hardware, software, and algorithms.[723] Hardware systems are divided into sensing and execution subsystems. The sensing unit captures real-time parameters through multi-sensor arrays, including spectroscopic, environmental, and visual inputs. The execution unit features automated reactors, microscale liquid handlers, and multi-degree-of-freedom robotic arms with sub-millimeter accuracy. Software architecture follows a layered design. A control layer, guided by real-time operating systems, coordinates hardware with precise, time-triggered tasks and built-in fault detection. Algorithmically, the emphasis is on actively exploring the experimental space. Active learning strategies select samples with high information value, balancing exploration and exploitation. Modern autonomous labs now embody a sense–decide–execute loop, autonomously designing experiments, analyzing data in real time, and dynamically optimizing research.

5.4.2. High-throughput experimentation accelerating combinatorial optimization. To optimize the crucial ion conductivity of liquid electrolytes, Krishnamoorthy *et al.*[724] proposed a specialized HTE platform to elucidate the effects of electrolyte composition on conductivity. The system fully automates the fast, systematic formulation of up to 96 distinct liquid electrolytes per working day, accommodating a broad range of Li salts, solvents/co-solvents, and multi-functional additives in both solid and liquid forms, all of which can be varied in composition and quantity. Extending this approach, Yan *et al.*[725] proposed a modular platform that integrates an automated HTE system with the liquid electrolyte component analysis (LECA) software package for data-driven modeling and analysis (Fig. 17a). Automated electrolyte formulation and conductivity measurements are performed by an HTE unit featuring both an automated preparation module and an automated conductivity testing module. The LECA software package streamlines data handling by incorporating widely used ML libraries into a simplified workflow that supports parallel training, cross-validation, and uncertainty estimation using linear regression, RF, neural networks, and GPR. By comparing prediction accuracy scores, LECA identifies the best-performing models and employs them to determine electrolyte compositions that maximize ionic conductivity under varying temperatures.

Further advancing the automation of electrolyte research, Noh *et al.*[726] introduced a fully autonomous workflow combining ML prediction with automated experiments, aiming to significantly enhance the solubility of redox-active organic molecules (Fig. 17b). Taking 2,1,3-benzothiadiazole as the model redox-active organic molecule, the researchers developed a closed-loop solvent screening process, composed of two interconnected modules: an HTE module and a BO module. The HTE module employs a high-throughput robotic platform for sample preparation and solubility measurements, while the BO module uses a surrogate model and acquisition function to predict solubility and recommend new solvents for evaluation. Experimental efficiency was significantly enhanced by the proposed workflow, which enabled the solubility of 42 samples to be measured within approximately 27 hours, over 13 times faster than the conventional approach (Fig. 17c). Human intervention was limited to transferring samples between the robotic system and the nuclear magnetic resonance instrument, thereby reducing the risk of operator-induced errors. Compared with random selection, the BO algorithm considerably accelerated the screening process by more effectively pinpointing solvents with high solubility. In the final screening of 2003 binary solvent systems, three rounds of BO identified 18 new mixtures in which 2,1,3-benzothiadiazole solubility exceeded 6.20 M, achieving a remarkable solubility of 6.50 M.

To generate comprehensive datasets and optimize the polymer electrolyte compositions, Stolberg *et al.*[727] proposed a fully automated characterization workflow to accelerate the discovery of polymer-based electrolytes (Fig. 17d). The high-throughput workflow involves defining experimental objectives, formulating recipes, establishing process parameters, and then employing ML for plan optimization (Fig. 17e). Subsequently, the raw materials (either commercially purchased or synthesized in the laboratory) and operating instructions were loaded into a high-throughput platform, by which all subsequent operations are automatically executed. The system was composed of modular workstations capable of chemical dispensing, mixing, drying, and physicochemical characterization, all executed by two three-axis robotic arms equipped with multifunctional tool heads. In terms of speed, the platform completes 90 sample tests within five days, which is around 100 times faster than traditional approaches. Moreover, the study yielded the largest dataset to date for comparing Li and sodium polymer electrolytes, encompassing over 70 unique formulations, 330 samples, and nearly 2000 ionic conductivity measurements. The dataset constitutes a valuable resource for guiding further research in polymer electrolytes.

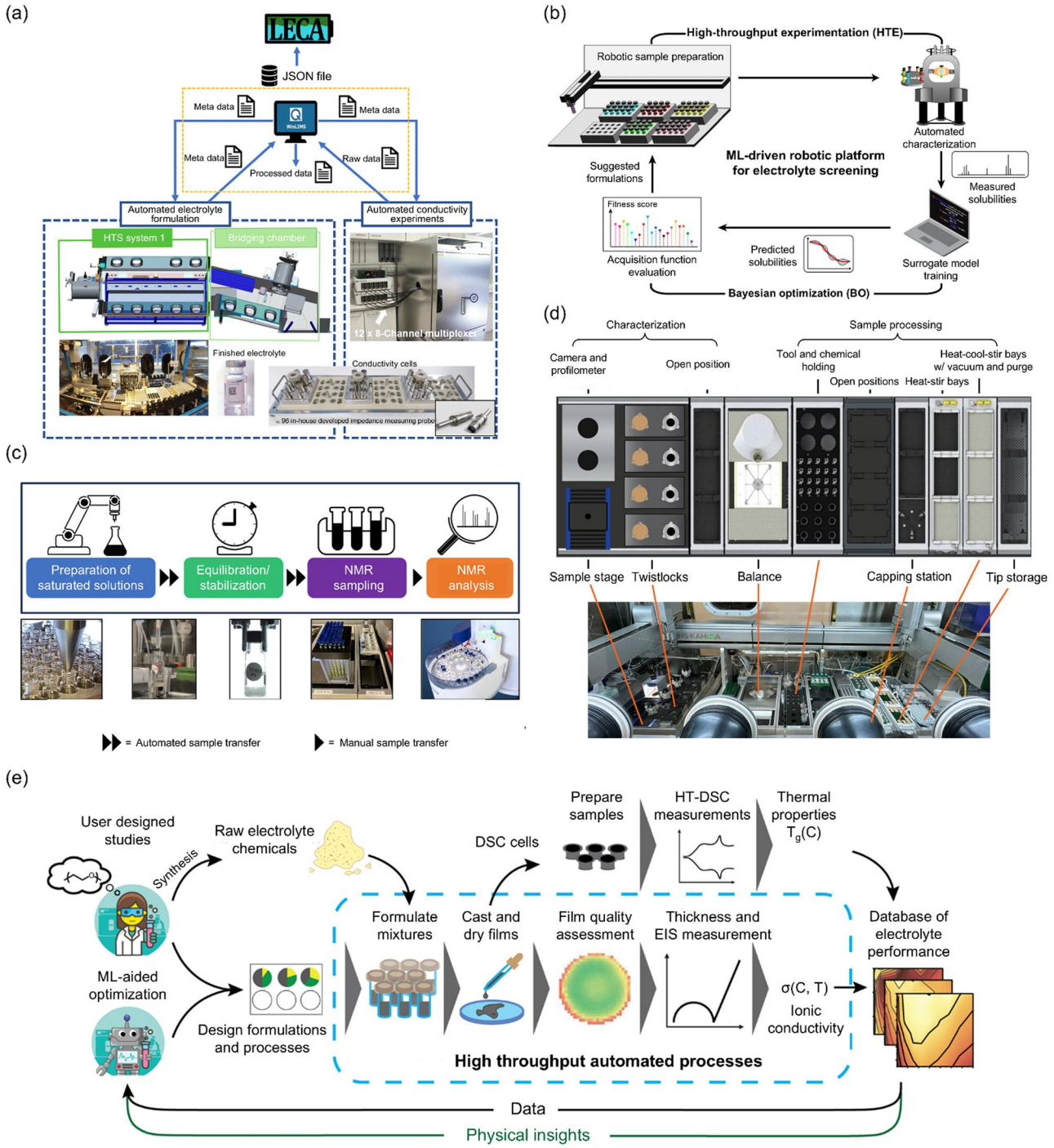

Fig. 17 High-throughput experimentation accelerating combinatorial optimization. (a) Non-aqueous electrolyte optimization combines HTE with ML to enhance ionic conductivity.[725] Reproduced with permission from ref. 725. Copyright 2024 The Royal Society of Chemistry. (b) A closed-loop ML-guided HTE platform integrates robotic sample preparation, automated characterization, and BO to iteratively screen electrolyte formulations. Experimental solubility data train surrogate models to prioritize candidates for subsequent rounds.[726] (c) Automated HTE workflow for solubility measurement executes powder/solvent dispensing, saturation monitoring, and NMR analysis.[726] Reproduced with permission from ref. 726. Copyright 2024 The Authors. (d) Polymer electrolyte characterization platform operates within a glovebox environment with modular deck components for automated testing.[727] (e) High-throughput polymer electrolyte evaluation outputs ionic conductivity *versus* salt concentration and temperature, supplemented by differential scanning calorimetry (DSC)-derived thermal properties.[727] Reproduced with permission from ref. 727. Copyright 2025 Elsevier Inc.

6. Conclusions and perspectives

The advancement of battery technologies hinges on the strategic design of functional molecules, which govern critical electrochemical processes across electrolytes, electrodes, and auxiliary components. Molecular innovation serves as the cornerstone for optimizing energy density, power density, cycling lifespan, working temperatures, and safety in next-generation battery systems. From redox-active organics enabling high-voltage cathodes to solvation-tailored electrolytes suppressing interfacial degradation, the discovery of novel molecules directly translates to transformative breakthroughs in energy storage.

AI has emerged as a paradigm-shifting tool to accelerate the molecular discovery process, bridging the gap between empirical exploration and rational design. By leveraging data-driven models, AI deciphers complex structure–property relationships at unprecedented speeds, enabling rapid prediction of interested properties such as redox potential, ionic conductivity, and viscosity. In addition, AI uncovers hidden patterns in chemical space, guiding the identification of molecular motifs that defy conventional design principles. Through virtual screening, generative design, and autonomous experimentation, AI transforms molecular discovery from a trial-and-error endeavor into a systematic, knowledge-driven discipline, positioning itself as an indispensable ally in the quest for sustainable energy storage solutions.

To fully realize the potential of AI in battery molecular innovation, concerted efforts should address four pivotal frontiers:

(1) Data infrastructure: establishing standardized and high-fidelity molecular databases is paramount for advancing AI-driven molecular design. Current datasets are often fragmented and heterogeneous, characterized by inconsistent experimental protocols, missing metadata, and limited representation of emerging chemistries such as SPE and multivalent ion systems. To address these limitations, the development of open-access and community-curated repositories is essential. The databases should integrate multi-scale data, spanning from QM calculations and MD simulations to macroscopic properties and device-level performance metrics. Such a comprehensive integration will facilitate the training of more generalizable and transferable AI models. Equally important is the implementation of robust data governance practices. Ensuring data quality, reproducibility, and interoperability requires the adoption of standardized ontologies, metadata schemas, and version control mechanisms. Moreover, the promotion of FAIR (findable, accessible, interoperable, and reusable) principles[728] will improve data discoverability and reusability across research domains. Federated learning frameworks offer a promising solution for harmonizing distributed and proprietary datasets, allowing institutions to collaboratively train models without sharing raw data, thus preserving confidentiality and enhancing inclusivity.[119,729] Ultimately, a coordinated effort in both data creation and governance will be vital for unlocking the full potential of AI in molecular discovery.

(2) Algorithmic synergy: future algorithm development should prioritize small-sample learning, transfer learning, and the design of algorithms tailored to complex systems. Battery research often suffers from data scarcity, making it essential to develop goal-oriented small-sample learning strategies. Advanced feature extraction methods are needed in this context. For example, architectures such as ChemXTree[730] and Meta-GAT[731] have been developed to enable the effective utilization of sparse data. Further model development is necessary, including both universal pre-trained models and dedicated few-shot learning frameworks. For instance, the pre-trained GIMLET model[732] enables zero-shot learning, while the SPARKLE framework[733] has been developed for zero-shot discovery of high-performance, low-cost organic battery materials. Supporting tools like knowledge graphs also show promise in addressing data scarcity. Methods such as KnowDDI[734] and EmerGNN[735] have been introduced to facilitate the efficient extraction of valuable information. Moreover, the rapid development of small-molecule algorithms in diverse fields such as drug discovery, perovskite batteries, and optoelectronic materials is expected to greatly contribute to the advancement of AI methods in battery research. These algorithms are often not confined to a single domain, and the emergence of general-purpose molecular models is likely to drive collective progress across the research community. In this process, developing advanced transfer learning techniques will be crucial. Integrating domain knowledge with modeling approaches may lead to the creation of highly specialized models better suited to the unique requirements of battery systems. Finally, since batteries represent inherently complex systems, establishing cross-scale AI models is a vital direction. Feng and co-workers[736–740] developed constant-potential methods to simulate the electrical double layers formed between the electrode and the electrolyte, thereby overcoming a challenge that could not be accurately addressed by conventional approaches and made a breakthrough in advancing the study of interfacial behavior. Cheng and co-workers[741–743] introduced a set of advanced hybrid algorithms for simulating electrode–electrolyte interfaces under operating conditions, offering new insights into interfacial structures and reactions. Future algorithm development should explicitly consider realistic scenarios such as multi-component electrolyte interactions and electrode–electrolyte interfacial reactions, while also extending molecular representation methods to capture this hierarchical complexity.

(3) Computational power: computational power forms the backbone of AI-driven molecular discovery, enabling large-scale model training and algorithm optimization for high-dimensional property prediction. Cloud-based platforms democratize access to GPU and tensor processing unit clusters, fostering collaborative workflows and accelerating virtual screening. In the future, quantum-classical hybrid architectures are expected to play a transformative role by combining the parallelism of quantum computing with the robustness of classical systems. This synergy holds the potential to overcome current computational bottlenecks, enabling real-time modeling of complex molecular systems and reaction networks. However, the energy demands of AI are escalating at an unprecedented rate. Recent estimates suggest that electricity consumption attributable to AI could surpass 1000 TWh by 2026, comparable to Japan's total annual consumption,

and rise to between 5% and 9% of global electricity usage by 2050 if the current trends continue.[744] This underscores the urgent need to develop energy-efficient computational frameworks. Balancing performance and efficiency will be essential to ensure that AI-driven molecular science progresses sustainably.

(4) Autonomous experimentation: closing the loop between AI predictions and experimental validation necessitates the widespread adoption of autonomous experimentation systems. Self-driving laboratories, integrating robotic synthesis platforms with real-time, high-resolution characterization tools, such as *operando* spectroscopy, high-throughput electrochemistry, and automated microscopy, will enable rapid and iterative design–test–learn cycles. Achieving these goals will require tight collaboration across disciplines, bringing together expertise in materials science, robotics, software engineering, and ML. Such integrated ecosystems will not only improve throughput and reproducibility but also enhance the adaptive learning capacity of AI models through real-time feedback loops. Ultimately, the fusion of AI-driven hypothesis generation with autonomous experimentation will usher in a new era of closed-loop scientific discovery in molecule and materials research.

The convergence of these pathways will redefine the molecular discovery landscape, fostering a new era where AI not only predicts but also understands and innovates. By harmonizing data, algorithms, computation, and experimentation, the scientific community can unlock the full potential of AI-driven molecular engineering, ultimately delivering battery systems that meet the urgent demands of a decarbonized energy future.

This is a beautiful era, for AI affords us the imaginative space to explore chemical phenomena and to construct advanced materials; it is also the worst of times, for AI that can truly deliver precisely as directed remains exceedingly rare. At the heart of these endeavours lies innovation—research conducted around genuine problems, with the aim of addressing real issues and genuinely solving them. A new age is dawning, in which data will emerge as a novel means of apprehending the world. High-level innovation will provide an unceasing stream of breakthroughs, and AI empowerment will drive energy chemistry and materials chemistry to become new quality productive forces for societal well-being and the advancement of a community with a shared future for mankind.

Conflicts of interest

There are no conflicts to declare.

Data availability

No primary research results, software, or code have been included, and no new data were generated or analysed as part of this review.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (T2322015, 22109007, 22393900, 22109011, 52394170, 52394171, 22109086, 22209010, 92472101 and 22108151), Beijing Municipal Natural Science Foundation (L233004 and L247015), and Tsinghua University Initiative Scientific Research Program. We thank Zheng Li, Si-Yuan Liu, Jin-Kai Tao, and Ming-Kang Liu for their helpful discussion.

References

1 International Energy Agency, Net Zero by 2050, International Energy Agency, Paris, 2024.
2 United Nations, The Paris Agreement, UN Climate Change Conference, Paris, 2015.
3 E. T. C. Vogt and B. M. Weckhuysen, *Nature*, 2024, **629**, 295–306.
4 International Energy Agency, World Energy Outlook 2024, International Energy Agency, Paris, 2024.
5 S. Chu and A. Majumdar, *Nature*, 2012, **488**, 294–303.
6 R. Teixeira, A. Cerveira, E. J. S. Pires and J. Baptista, *Energies*, 2024, **17**, 3480.
7 International Energy Agency, Batteries and Secure Energy Transitions, International Energy Agency, Paris, 2024.
8 B. Dunn, H. Kamath and J.-M. Tarascon, *Science*, 2011, **334**, 928–935.
9 A. Celadon, H. Sun, S. Sun and G. Zhang, *SusMat*, 2024, **4**, e234.
10 C. Shao, Y. Zhao and L. Qu, *SusMat*, 2022, **2**, 142–160.
11 D. Lu, R. Li, M. M. Rahman, P. Yu, L. Lv, S. Yang, Y. Huang, C. Sun, S. Zhang, H. Zhang, J. Zhang, X. Xiao, T. Deng, L. Fan, L. Chen, J. Wang, E. Hu, C. Wang and X. Fan, *Nature*, 2024, **627**, 101–107.
12 J. Xie, Z. Liang and Y.-C. Lu, *Nat. Mater.*, 2020, **19**, 1006–1011.
13 B. Ma, H. Zhang, R. Li, S. Zhang, L. Chen, T. Zhou, J. Wang, R. Zhang, S. Ding, X. Xiao, T. Deng, L. Chen and X. Fan, *Nat. Chem.*, 2024, **16**, 1427–1435.
14 Y.-X. Yao, L. Xu, C. Yan and Q. Zhang, *EES Batteries*, 2025, **1**, 9–22.
15 S.-Y. Sun, X.-Q. Zhang, X.-Y. Yan, Z. Zheng, Q.-K. Zhang and J.-Q. Huang, *EES Batteries*, 2025, **1**, 340–363.
16 A. Yang, X. Gao, M. Pei, J. Zhou, H. Wang, C. Liao, J. Xiao, Y. Liu, W. Yan and J. Zhang, *Angew. Chem., Int. Ed.*, 2025, **64**, e202424237.
17 Y. S. Meng, V. Srinivasan and K. Xu, *Science*, 2022, **378**, eabq3750.
18 Z. Li, R. Yu, S. Weng, Q. Zhang, X. Wang and X. Guo, *Nat. Commun.*, 2023, **14**, 482.
19 A. Szczęsna-Chrzan, M. Vogler, P. Yan, G. Z. Żukowska, C. Wölke, A. Ostrowska, S. Szymańska, M. Marcinek, M. Winter, I. Cekic-Laskovic, W. Wieczorek and H. S. Stein, *J. Mater. Chem. A*, 2023, **11**, 13483–13492.
20 O. Borodin, X. Ren, J. Vatamanu, A. von Wald Cresce, J. Knap and K. Xu, *Acc. Chem. Res.*, 2017, **50**, 2886–2894.
21 C. F. N. Marchiori, R. P. Carvalho, M. Ebadi, D. Brandell and C. M. Araujo, *Chem. Mater.*, 2020, **32**, 7237–7246.
22 X. Fan, X. Ji, L. Chen, J. Chen, T. Deng, F. Han, J. Yue, N. Piao, R. Wang, X. Zhou, X. Xiao, L. Chen and C. Wang, *Nat. Energy*, 2019, **4**, 882–890.

23 Y. Ou, P. Zhou, W. Hou, X. Ma, X. Song, S. Yan, Y. Lu and K. Liu, *J. Energy Chem.*, 2024, **94**, 360–392.
24 Z. Li, B.-Q. Li, C.-X. Bi, X.-Y. Li, M. Zhao and Q. Zhang, *Mater. Sci. Eng. R Rep.*, 2025, **164**, 100955.
25 Z. Li, Y. Li, C.-X. Bi, Q.-K. Zhang, L.-P. Hou, X.-Y. Li, J. Ma, X.-Q. Zhang, B.-Q. Li, R. Wen and Q. Zhang, *Adv. Funct. Mater.*, 2024, **34**, 2304541.
26 M. Zhao, X.-Y. Li, X. Chen, B.-Q. Li, S. Kaskel, Q. Zhang and J.-Q. Huang, *eScience*, 2021, **1**, 44–52.
27 Q. Cheng, Z.-X. Chen, X.-Y. Li, L.-P. Hou, C.-X. Bi, X.-Q. Zhang, J.-Q. Huang and B.-Q. Li, *J. Energy Chem*, 2023, **76**, 181–186.
28 L.-P. Hou, L.-Y. Yao, C.-X. Bi, J. Xie, B.-Q. Li, J.-Q. Huang and X.-Q. Zhang, *J. Energy Chem.*, 2022, **68**, 300–305.
29 L.-P. Hou, X.-Q. Zhang, N. Yao, X. Chen, B.-Q. Li, P. Shi, C.-B. Jin, J.-Q. Huang and Q. Zhang, *Chem*, 2022, **8**, 1083–1098.
30 L. P. Hou, Z. Li, N. Yao, C. X. Bi, B. Q. Li, X. Chen, X. Q. Zhang and Q. Zhang, *Adv. Mater.*, 2022, **34**, 2205284.
31 L.-L. Su, N. Yao, Z. Li, C.-X. Bi, Z.-X. Chen, X. Chen, B.-Q. Li, X.-Q. Zhang and J.-Q. Huang, *Angew. Chem., Int. Ed.*, 2024, **63**, e202318785.
32 X.-Y. Li, M. Zhao, Y.-W. Song, C.-X. Bi, Z. Li, Z.-X. Chen, X.-Q. Zhang, B.-Q. Li and J.-Q. Huang, *Chem. Soc. Rev.*, 2025, **54**, 4822–4873.
33 Z. Li, L.-P. Hou, N. Yao, X.-Y. Li, Z.-X. Chen, X. Chen, X.-Q. Zhang, B.-Q. Li and Q. Zhang, *Angew. Chem., Int. Ed.*, 2023, **62**, e202309968.
34 J.-N. Liu, C.-X. Zhao, J. Wang, D. Ren, B.-Q. Li and Q. Zhang, *Energy Environ. Sci.*, 2022, **15**, 4542–4553.
35 C.-X. Zhao, J.-N. Liu, J. Wang, D. Ren, B.-Q. Li and Q. Zhang, *Chem. Soc. Rev.*, 2021, **50**, 7745–7778.
36 C.-X. Zhao, L. Yu, J.-N. Liu, J. Wang, N. Yao, X.-Y. Li, X. Chen, B.-Q. Li and Q. Zhang, *Angew. Chem., Int. Ed.*, 2022, **61**, e202208042.
37 H. Fu, S. Huang, C. Wang, J. S. Kim, Y. Zhao, Y. Wu, P. Xiong and H. S. Park, *Adv. Energy Mater.*, 2025, **15**, 2501152.
38 S. Huang, P. Zhang, J. Lu, J. S. Kim, D. H. Min, J. S. Byun, M. J. Kim, H. Fu, P. Xiong, P. J. Yoo, W. Li, X. Yu, X. Qin and H. S. Park, *Energy Environ. Sci.*, 2024, **17**, 7870–7881.
39 Q. Zhao, S. Stalin, C.-Z. Zhao and L. A. Archer, *Nat. Rev. Mater.*, 2020, **5**, 229–252.
40 Q. Zhou, J. Ma, S. Dong, X. Li and G. Cui, *Adv. Mater.*, 2019, **31**, 1902029.
41 P. Xu, Y.-C. Gao, Y.-X. Huang, Z.-Y. Shuang, W.-J. Kong, X.-Y. Huang, W.-Z. Huang, N. Yao, X. Chen, H. Yuan, C.-Z. Zhao, J.-Q. Huang and Q. Zhang, *Adv. Mater.*, 2024, **36**, 2409489.
42 J.-K. Hu, Y.-C. Gao, S.-J. Yang, X.-L. Wang, X. Chen, Y.-L. Liao, S. Li, J. Liu, H. Yuan and J.-Q. Huang, *Adv. Funct. Mater.*, 2024, **34**, 2311633.
43 K. Amini, A. N. Shocron, M. E. Suss and M. J. Aziz, *ACS Energy Lett.*, 2023, **8**, 3526–3535.
44 Z. Yu, X. Jia, Y. Cai, R. Su, Q. Zhu, T. Zhao and H. Jiang, *Energy Storage Mater.*, 2024, **69**, 103404.
45 S. Ahn, A. Yun, D. Ko, V. Singh, J. M. Joo and H. R. Byon, *Chem. Soc. Rev.*, 2025, **54**, 742–789.
46 Y. Lu and J. Chen, *Nat. Rev. Chem.*, 2020, **4**, 127–142.
47 B. Esser, F. Dolhem, M. Becuwe, P. Poizot, A. Vlad and D. Brandell, *J. Power Sources*, 2021, **482**, 228814.
48 Y. Liu, Y.-B. Ma, W. Jaegermann, R. Hausbrand and B.-X. Xu, *J. Power Sources*, 2020, **454**, 227892.
49 Y. Liu, W.-B. Yu and B.-X. Xu, *J. Power Sources*, 2022, **534**, 231406.
50 Y. Lu, Y. Cai, Q. Zhang and J. Chen, *Adv. Mater.*, 2022, **34**, 2104150.
51 J. Heiska, M. Nisula and M. Karppinen, *J. Mater. Chem. A*, 2019, **7**, 18735–18758.
52 J. J. Shea and C. Luo, *ACS Appl. Mater. Interfaces*, 2020, **12**, 5361–5380.
53 Y. Liang, Z. Tao and J. Chen, *Adv. Energy Mater.*, 2012, **2**, 742–769.
54 Y. Liang, P. Zhang, S. Yang, Z. Tao and J. Chen, *Adv. Energy Mater.*, 2013, **3**, 600–605.
55 Y. Shirota and H. Kageyama, *Chem. Rev.*, 2007, **107**, 953–1010.
56 X. Zhao and X. Zhan, *Chem. Soc. Rev.*, 2011, **40**, 3728–3743.
57 Q. Zhao, Y. Lu and J. Chen, *Adv. Energy Mater.*, 2017, **7**, 1601792.
58 M. T. Jeena, J.-I. Lee, S. H. Kim, C. Kim, J.-Y. Kim, S. Park and J.-H. Ryu, *ACS Appl. Mater. Interfaces*, 2014, **6**, 18001–18007.
59 F. Zou and A. Manthiram, *Adv. Energy Mater.*, 2020, **10**, 2002508.
60 T. Qin, H. Yang, Q. Li, X. Yu and H. Li, *Ind. Chem. Mater.*, 2024, **2**, 191–225.
61 Y.-K. Hong, J.-H. Kim, N.-Y. Kim, K.-S. Oh, H.-I. Kim, S. Ryu, Y. Ko, J.-Y. Kim, K.-H. Lee and S.-Y. Lee, *Nano-Micro Lett.*, 2025, **17**, 112.
62 J.-H. Kim, K. M. Lee, J. W. Kim, S. H. Kweon, H.-S. Moon, T. Yim, S. K. Kwak and S.-Y. Lee, *Nat. Commun.*, 2023, **14**, 5721.
63 X. Zhang, Y. Wu, B. Yu, K. Hu, P. Zhang, F. Ding, L. Zhang, Y. Chen, J. Z. Ou and Z. Zhang, *EcoEnergy*, 2024, **2**, 549–598.
64 S. Cao, J. Tan, L. Ma, Y. Liu, Q. He, W. Lu, Z. Liu, M. Ye and J. Shen, *Energy Storage Mater.*, 2024, **66**, 103232.
65 A. Benayad, D. Diddens, A. Heuer, A. N. Krishnamoorthy, M. Maiti, F. L. Cras, M. Legallais, F. Rahmanian, Y. Shin, H. Stein, M. Winter, C. Wölke, P. Yan and I. Cekic-Laskovic, *Adv. Energy Mater.*, 2022, **12**, 2102678.
66 N. Yao, X. Chen, Z. H. Fu and Q. Zhang, *Chem. Rev.*, 2022, **122**, 10970–11021.
67 L. Yu, X. Chen, N. Yao, Y.-C. Gao, Y.-H. Yuan, Y.-B. Gao, C. Tang and Q. Zhang, *InfoMat*, 2025, **7**, e12653.
68 L. Yu, X. Chen, N. Yao, Y.-C. Gao and Q. Zhang, *J. Mater. Chem. A*, 2023, **11**, 11078–11088.
69 L. Yu, N. Yao, Y.-C. Gao, Z.-H. Fu, B. Jiang, R. Li, C. Tang and X. Chen, *J. Energy Chem.*, 2024, **93**, 299–305.
70 R. Zhang, X. Shen, Y.-T. Zhang, X.-L. Zhong, H.-T. Ju, T.-X. Huang, X. Chen, J.-D. Zhang and J.-Q. Huang, *J. Energy Chem.*, 2022, **71**, 29–35.
71 J.-L. Li, L. Shen, Z.-N. Cheng, J.-D. Zhang, L.-X. Li, Y.-T. Zhang, Y.-B. Gao, C. Guo, X. Chen, C.-Z. Zhao, R. Zhang and Q. Zhang, *J. Energy Chem.*, 2025, **101**, 16–22.

72 X. Chen, X. Liu, X. Shen and Q. Zhang, *Angew. Chem., Int. Ed.*, 2021, **60**, 24354–24366.
73 X. Chen, N. Yao, Z. Zheng, Y.-C. Gao and Q. Zhang, *Nat. Sci. Rev.*, 2024, **12**, nwae394.
74 K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, *Nature*, 2018, **559**, 547–555.
75 B. Sanchez-Lengeling and A. Aspuru-Guzik, *Science*, 2018, **361**, 360–365.
76 AI pioneers win 2024 Nobel prizes, *Nat. Mach. Intell.*, 2024, **6**, 1271.
77 N. L. Rider and M. Shamji, *J. Allergy Clin. Immunol.*, 2025, **155**, 808–809.
78 K. J. Kanarik, W. T. Osowiecki, Y. Lu, D. Talukder, N. Roschewsky, S. N. Park, M. Kamon, D. M. Fried and R. A. Gottscho, *Nature*, 2023, **616**, 707–711.
79 J. Wu, L. Torresi, M. Hu, P. Reiser, J. Zhang, J. S. Rocha-Ortiz, L. Wang, Z. Xie, K. Zhang, B. W. Park, A. Barabash, Y. Zhao, J. Luo, Y. Wang, L. Lüer, L. L. Deng, J. A. Hauch, D. M. Guldi, M. E. Pérez-Ojeda, S. I. Seok, P. Friederich and C. J. Brabec, *Science*, 2024, **386**, 1256–1264.
80 C. Zeni, R. Pinsler, D. Zügner, A. Fowler, M. Horton, X. Fu, Z. Wang, A. Shysheya, J. Crabbé, S. Ueda, R. Sordillo, L. Sun, J. Smith, B. Nguyen, H. Schulz, S. Lewis, C.-W. Huang, Z. Lu, Y. Zhou, H. Yang, H. Hao, J. Li, C. Yang, W. Li, R. Tomioka and T. Xie, *Nature*, 2025, **639**, 624–632.
81 T. Lombardo, M. Duquesnoy, H. El-Bouysidy, F. Årén, A. Gallo-Bueno, P. B. Jørgensen, A. Bhowmik, A. Demortière, E. Ayerbe, F. Alcaide, M. Reynaud, J. Carrasco, A. Grimaud, C. Zhang, T. Vegge, P. Johansson and A. A. Franco, *Chem. Rev.*, 2022, **122**, 10899–10969.
82 M. Aykol, P. Herring and A. Anapolsky, *Nat. Rev. Mater.*, 2020, **5**, 725–727.
83 Y. Liu, B. Guo, X. Zou, Y. Li and S. Shi, *Energy Storage Mater.*, 2020, **31**, 434–450.
84 Z. Wei, Q. He and Y. Zhao, *J. Power Sources*, 2022, **549**, 232125.
85 Y.-C. Gao, N. Yao, X. Chen, L. Yu, R. Zhang and Q. Zhang, *J. Am. Chem. Soc.*, 2023, **145**, 23764–23770.
86 Y.-C. Gao, Y.-H. Yuan, S. Huang, N. Yao, L. Yu, Y.-P. Chen, Q. Zhang and X. Chen, *Angew. Chem., Int. Ed.*, 2025, **64**, e202416506.
87 C. Han, Y.-C. Gao, X. Chen, X. Liu, N. Yao, L. Yu, L. Kong and Q. Zhang, *InfoMat*, 2024, **6**, e12521.
88 S. C. Kim, S. T. Oyakhire, C. Athanitis, J. Wang, Z. Zhang, W. Zhang, D. T. Boyle, M. S. Kim, Z. Yu, X. Gao, T. Sogade, E. Wu, J. Qin, Z. Bao, S. F. Bent and Y. Cui, *Proc. Natl. Acad. Sci. U. S. A.*, 2023, **120**, e2214357120.
89 H. Chen, M. Armand, G. Demailly, F. Dolhem, P. Poizot and J.-M. Tarascon, *ChemSusChem*, 2008, **1**, 348–355.
90 L. Cheng, R. S. Assary, X. Qu, A. Jain, S. P. Ong, N. N. Rajput, K. Persson and L. A. Curtiss, *J. Phys. Chem. Lett.*, 2015, **6**, 283–291.
91 K. Lin, Q. Chen, M. R. Gerhardt, L. Tong, S. B. Kim, L. Eisenach, A. W. Valle, D. Hardee, R. G. Gordon, M. J. Aziz and M. P. Marshak, *Science*, 2015, **349**, 1529–1532.
92 L. Suo, O. Borodin, T. Gao, M. Olguin, J. Ho, X. Fan, C. Luo, C. Wang and K. Xu, *Science*, 2015, **350**, 938–943.
93 Y.-X. Yao, X. Chen, C. Yan, X.-Q. Zhang, W.-L. Cai, J.-Q. Huang and Q. Zhang, *Angew. Chem., Int. Ed.*, 2021, **60**, 4090–4097.
94 J. H. Fletcher, *J. Chem. Doc.*, 1967, **7**, 64–67.
95 A. M. Patterson and C. E. Curran, *J. Am. Chem. Soc.*, 1917, **39**, 1623–1638.
96 S. Skonieczny, *J. Chem. Educ.*, 2006, **83**, 1633.
97 S. Raghunathan and U. D. Priyakumar, *Int. J. Quantum. Chem.*, 2022, **122**, e26870.
98 L. David, A. Thakkar, R. Mercado and O. Engkvist, *J. Cheminf.*, 2020, **12**, 56.
99 W. L. Chen, *J. Chem Inf. Model.*, 2006, **46**, 2230–2255.
100 A. Dietz, *J. Chem. Inf. Comp. Sci.*, 1995, **35**, 787–802.
101 K. V. Chuang, L. M. Gunsalus and M. J. Keiser, *J. Med. Chem.*, 2020, **63**, 8705–8722.
102 T.-H. Nguyen-Vo, P. Teesdale-Spittle, J. E. Harvey and B. P. Nguyen, *Memet. Comput.*, 2024, **16**, 519–536.
103 Y. Harnik and A. Milo, *Chem. Sci.*, 2024, **15**, 5052–5055.
104 M. McGibbon, S. Shave, J. Dong, Y. Gao, D. R. Houston, J. Xie, Y. Yang, P. Schwaller and V. Blay, *Brief. Bioinform.*, 2024, **25**, bbad422.
105 M. V. Sabando, I. Ponzoni, E. E. Milios and A. J. Soto, *Brief. Bioinform.*, 2022, **23**, bbab365.
106 Y. Li, T. Li and H. Liu, *Knowl. Inf. Syst.*, 2017, **53**, 551–577.
107 J. Cai, J. Luo, S. Wang and S. Yang, *Neurocomputing*, 2018, **300**, 70–79.
108 A. Y.-T. Wang, R. J. Murdock, S. K. Kauwe, A. O. Oliynyk, A. Gurlo, J. Brgoch, K. A. Persson and T. D. Sparks, *Chem. Mater.*, 2020, **32**, 4954–4965.
109 Z. Li, M. Jiang, S. Wang and S. Zhang, *Drug Discov. Today*, 2022, **27**, 103373.
110 K. Atz, F. Grisoni and G. Schneider, *Nat. Mach. Intell.*, 2021, **3**, 1023–1032.
111 Y. Du, X. Guo, Y. Wang, A. Shehu and L. Zhao, *Bioinformatics*, 2022, **38**, 3200–3208.
112 X. Fang, L. Liu, J. Lei, D. He, S. Zhang, J. Zhou, F. Wang, H. Wu and H. Wang, *Nat. Mach. Intell.*, 2022, **4**, 127–134.
113 R. Winter, F. Montanari, F. Noé and D.-A. Clevert, *Chem. Sci.*, 2019, **10**, 1692–1701.
114 A. L. Dias, L. Bustillo and T. Rodrigues, *Nat. Commun.*, 2023, **14**, 6394.
115 H. Li, R. Zhang, Y. Min, D. Ma, D. Zhao and J. Zeng, *Nat. Commun.*, 2023, **14**, 7568.
116 O. A. von Lilienfeld, R. Ramakrishnan, M. Rupp and A. Knoll, *Int. J. Quantum. Chem.*, 2015, **115**, 1084–1093.
117 Y. Bengio, A. Courville and P. Vincent, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013, **35**, 1798–1828.
118 R. A. Sayle, *J. Comput. Aid. Mol. Des*, 2010, **24**, 485–496.
119 J. Shen and C. A. Nicolaou, *Drug Discovery Today: Technologies*, 2019, **32–33**, 29–36.
120 M. Shahlaei, *Chem. Rev.*, 2013, **113**, 8093–8103.
121 H. Li, C. W. Yap, C. Y. Ung, Y. Xue, Z. W. Cao and Y. Z. Chen, *J. Chem Inf. Model.*, 2005, **45**, 1376–1384.
122 Y. Okamoto and Y. Kubo, *ACS Omega*, 2018, **3**, 7868–7874.

123 J. Deng, Z. Yang, H. Wang, I. Ojima, D. Samaras and F. Wang, *Nat. Commun.*, 2023, **14**, 6395.
124 V. c G. Satorras, E. Hoogeboom and M. Welling, in Proceedings of the 38th International Conference on Machine Learning, 2021.
125 J. Fei and Z. Deng, *Artif. Intell. Rev.*, 2024, **57**, 168.
126 S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt and B. Kozinsky, *Nat. Commun.*, 2022, **13**, 2453.
127 Y. Wang, T. Wang, S. Li, X. He, M. Li, Z. Wang, N. Zheng, B. Shao and T.-Y. Liu, *Nat. Commun.*, 2024, **15**, 313.
128 I. Igashov, H. Stärk, C. Vignac, A. Schneuing, V. G. Satorras, P. Frossard, M. Welling, M. Bronstein and B. Correia, *Nat. Mach. Intell.*, 2024, **6**, 417–427.
129 F. Noé, A. Tkatchenko, K.-R. Müller and C. Clementi, *Annu. Rev. Phys. Chem.*, 2020, **71**, 361–390.
130 S. Villar, D. W. Hogg, K. Storey-Fisher, W. Yao and B. Blum-Smith, in *Advances in Neural Information Processing Systems*, 2021.
131 C. Isert, K. Atz and G. Schneider, *Curr. Opin. Struct. Biol.*, 2023, **79**, 102548.
132 A. Dumitrescu, D. Korpela, M. Heinonen, Y. Verma, V. Iakovlev, V. Garg and H. Lähdesmäki, in *The Thirteenth International Conference on Learning Representations*, 2025.
133 A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu and D. Hassabis, *Nature*, 2020, **577**, 706–710.
134 J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, *Nature*, 2021, **596**, 583–589.
135 J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis and J. M. Jumper, *Nature*, 2024, **630**, 493–500.
136 G. Zhou, Z. Gao, Q. Ding, H. Zheng, H. Xu, Z. Wei, L. Zhang and G. Ke, in *The Eleventh International Conference on Learning Representations*, 2023.
137 X. Ji, Z. Wang, Z. Gao, H. Zheng, L. Zhang and G. Ke, in *Advances in Neural Information Processing Systems*, 2024, vol. 37.
138 X. Li and D. Fourches, *J. Chem Inf. Model.*, 2021, **61**, 1560–1569.
139 B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams and A. G. Doyle, *Nature*, 2021, **590**, 89–96.
140 E. Poslavskaya and A. Korolev, *arXiv*, 2023, preprint, arXiv:2312.16930, DOI: **10.48550/arXiv.2312.16930**.
141 M. K. Dahouda and I. Joe, *IEEE Access*, 2021, **9**, 114381–114391.
142 B. Ranković, R.-R. Griffiths, H. B. Moss and P. Schwaller, *Digit. Discov.*, 2024, **3**, 654–666.
143 A. Pomberger, A. A. Pedrina McCarthy, A. Khan, S. Sung, C. J. Taylor, M. J. Gaunt, L. Colwell, D. Walz and A. A. Lapkin, *React. Chem. Eng.*, 2022, **7**, 1368–1379.
144 E. Nuñez-Andrade, I. Vidal-Daza, J. W. Ryan, R. Gómez-Bombarelli and F. J. Martin-Martinez, *Digit. Discov.*, 2025, **4**, 776–789.
145 J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals and G. E. Dahl, in *Proceedings of the 34th International Conference on Machine Learning*, 2017.
146 W. J. Wiswesser, *Chem. Eng. News*, 1952, **30**, 3523–3526.
147 W. J. Wiswesser, *J. Chem. Doc.*, 1968, **8**, 146–150.
148 W. J. Wiswesser, *J. Chem. Inf. Comp. Sci.*, 1982, **22**, 88–93.
149 J. J. Vollmer, *J. Chem. Educ.*, 1983, **60**, 192.
150 D. Weininger, *J. Chem. Inf. Comp. Sci.*, 1988, **28**, 31–36.
151 E. J. Bjerrum and B. Sattarov, *Biomolecules*, 2018, **8**, 131.
152 M. Moret, L. Friedrich, F. Grisoni, D. Merk and G. Schneider, *Nat. Mach. Intell.*, 2020, **2**, 171–180.
153 L. Cui, H. Li, K. Chen, L. Shou and G. Chen, *arXiv*, 2024, preprint, arXiv:2407.21523, DOI: **10.48550/arXiv.2407.21523**.
154 A. Gangwal, A. Ansari, I. Ahmad, A. K. Azad and W. M. A. Wan Sulaiman, *Comput. Biol. Med.*, 2024, **179**, 108734.
155 J. Arús-Pous, S. V. Johansson, O. Prykhodko, E. J. Bjerrum, C. Tyrchan, J.-L. Reymond, H. Chen and O. Engkvist, *J. Cheminf.*, 2019, **11**, 71.
156 H. L. Morgan, *J. Chem. Doc.*, 1965, **5**, 107–113.
157 D. Weininger, A. Weininger and J. L. Weininger, *J. Chem. Inf. Comp. Sci.*, 1989, **29**, 97–101.
158 N. M. O'Boyle, *J. Cheminf.*, 2012, **4**, 22.
159 N. Schneider, R. A. Sayle and G. A. Landrum, *J. Chem Inf. Model.*, 2015, **55**, 2111–2120.
160 D. G. Krotko, *J. Cheminf.*, 2020, **12**, 48.
161 M. Krenn, F. Häse, A. Nigam, P. Friederich and A. Aspuru-Guzik, *Mach. Learn.: Sci. Technol*, 2020, **1**, 045024.
162 M. Krenn, Q. Ai, S. Barthel, N. Carson, A. Frei, N. C. Frey, P. Friederich, T. Gaudin, A. A. Gayle, K. M. Jablonka, R. F. Lameiro, D. Lemm, A. Lo, S. M. Moosavi, J. M. Nápoles-Duarte, A. Nigam, R. Pollice, K. Rajan, U. Schatzschneider, P. Schwaller, M. Skreta, B. Smit, F. Strieth-Kalthoff, C. Sun, G. Tom, G. Falk von Rudorff, A. Wang, A. D. White, A. Young, R. Yu and A. Aspuru-Guzik, *Patterns*, 2022, **3**, 100588.
163 N. O'Boyle and A. Dalke, *ChemRxiv*, 2018, preprint, DOI: **10.26434/chemrxiv.7097960.v1**.
164 K. Rajan, C. Steinbeck and A. Zielesny, *Digit. Discov.*, 2022, **1**, 84–90.

165 A language for describing molecular patterns, https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html, accessed 20 March 2025.
166 R. Schmidt, E. S. R. Ehmki, F. Ohm, H.-C. Ehrlich, A. Mashychev and M. Rarey, *J. Chem. Inf. Model.*, 2019, **59**, 2560–2571.
167 E. S. R. Ehmki, R. Schmidt, F. Ohm and M. Rarey, *J. Chem. Inf. Model.*, 2019, **59**, 2572–2586.
168 R. K. Harris, E. D. Becker, S. M. Cabral de Menezes, R. Goodfellow and P. Granger, *Magn. Reson. Chem.*, 2002, **40**, 489–505.
169 D. Everett and L. Koopal, *Polymer*, 2001, **31**, 1598.
170 G. J. Leigh, *Principles of chemical nomenclature: a guide to IUPAC recommendations*, Royal Society of Chemistry, Cambridge, 2011.
171 P. G. Dittmar, N. A. Farmer, W. Fisanick, R. C. Haines and J. Mockus, *J. Chem. Inf. Comp. Sci.*, 1983, **23**, 93–102.
172 D. W. Weisgerber, *J. Am. Soc. Inf. Sci.*, 1997, **48**, 349–360.
173 S. Heller, A. McNaught, S. Stein, D. Tchekhovskoi and I. Pletnev, *J. Cheminf.*, 2013, **5**, 7.
174 G. Grethe, J. M. Goodman and C. H. G. Allen, *J. Cheminf.*, 2013, **5**, 45.
175 S. R. Heller, A. McNaught, I. Pletnev, S. Stein and D. Tchekhovskoi, *J. Cheminf.*, 2015, **7**, 23.
176 I. Pletnev, A. Erin, A. McNaught, K. Blinov, D. Tchekhovskoi and S. Heller, *J. Cheminf.*, 2012, **4**, 39.
177 C. Southan, *J. Cheminf.*, 2013, **5**, 10.
178 S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang and E. E. Bolton, *Nucleic. Acids. Res.*, 2021, **49**, D1388–D1395.
179 H. E. Pence and A. Williams, *J. Chem. Educ.*, 2010, **87**, 1123–1124.
180 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, *APL Mater.*, 2013, **1**, 011002.
181 A. Jain, J. Montoya, S. Dwaraknath, N. E. R. Zimmermann, J. Dagdelen, M. Horton, P. Huck, D. Winston, S. Cholia, S. P. Ong and K. Persson, *The Materials Project: Accelerating Materials Design Through Theory-Driven Data and Tools*, Springer International Publishing, Cham, 2020.
182 L. Pattanaik and C. W. Coley, *Chem*, 2020, **6**, 1204–1207.
183 J. Yang, Y. Cai, K. Zhao, H. Xie and X. Chen, *Drug Discov. Today*, 2022, **27**, 103356.
184 B. D. Christie, B. A. Leland and J. G. Nourse, *J. Chem. Inf. Comp. Sci.*, 1993, **33**, 545–547.
185 D. Boldini, D. Ballabio, V. Consonni, R. Todeschini, F. Grisoni and S. A. Sieber, *J. Cheminf.*, 2024, **16**, 35.
186 K.-Z. Myint, L. Wang, Q. Tong and X.-Q. Xie, *Mol. Pharm.*, 2012, **9**, 2912–2923.
187 L. Xie, L. Xu, R. Kong, S. Chang and X. Xu, *Front. Pharmacol.*, 2020, **11**, 606668.
188 J. L. Durant, B. A. Leland, D. R. Henry and J. G. Nourse, *J. Chem. Inf. Comp. Sci.*, 2002, **42**, 1273–1280.
189 RDKit: Open-source cheminformatics. https://www.rdkit.org/, accessed 25 April 2025.
190 N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, *J. Cheminf.*, 2011, **3**, 33.
191 E. E. Bolton, Y. Wang, P. A. Thiessen and S. H. Bryant, *Annual Reports in Computational Chemistry*, Elsevier, Amsterdam, 2008.
192 D. J. Rogers and T. T. Tanimoto, *Science*, 1960, **132**, 1115–1118.
193 A. Cereto-Massagué, M. J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé and G. Pujadas, *Methods*, 2015, **71**, 58–63.
194 D. Rogers and M. Hahn, *J. Chem Inf. Model.*, 2010, **50**, 742–754.
195 M. Hassan, R. D. Brown, S. Varma-O'Brien and D. Rogers, *Mol. Divers.*, 2006, **10**, 283–299.
196 Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang and S. H. Bryant, *Nucleic. Acids. Res.*, 2009, **37**, W623–W633.
197 J. Schwartz, M. Awale and J.-L. Reymond, *J. Chem Inf. Model.*, 2013, **53**, 1979–1989.
198 N. M. O'Boyle and R. A. Sayle, *J. Cheminf.*, 2016, **8**, 36.
199 D. Probst and J. L. Reymond, *J. Cheminf.*, 2018, **10**, 66.
200 A. Bender, H. Y. Mussa, R. C. Glen and S. Reiling, *J. Chem. Inf. Comp. Sci.*, 2004, **44**, 1708–1718.
201 A. Bender, H. Y. Mussa, G. S. Gill and R. C. Glen, *J. Med. Chem.*, 2004, **47**, 6569–6583.
202 R. E. Carhart, D. H. Smith and R. Venkataraghavan, *J. Chem. Inf. Comp. Sci*, 1985, **25**, 64–73.
203 M. Awale and J.-L. Reymond, *J. Chem Inf. Model.*, 2014, **54**, 1892–1907.
204 R. Nilakantan, N. Bauman and J. S. Dixon, *J. Chem. Inf. Comput. Sci.*, 1987, **27**, 82–85.
205 P. C. D. Hawkins, A. G. Skillman and A. Nicholls, *J. Med. Chem.*, 2007, **50**, 74–82.
206 P. J. Ballester and W. G. Richards, *J. Comput. Chem.*, 2007, **28**, 1711–1723.
207 A. T. Balaban, *J. Chem. Inf. Comp. Sci.*, 1985, **25**, 334–343.
208 J. M. Amigó, J. Gálvez and V. M. Villar, *Naturwissenschaften*, 2009, **96**, 749–761.
209 J. A. Bondy and U. S. R. Murty, *Graph theory with applications*, Macmillan, London, 1976.
210 J. L. Gross, J. Yellen and M. Anderson, *Graph theory and its applications*, Chapman & Hall, New York, 2018.
211 R. García-Domenech, J. Gálvez, J. V. de Julián-Ortiz and L. Pogliani, *Chem. Rev.*, 2008, **108**, 1127–1169.
212 L. S. G. Leite, S. Banerjee, Y. Wei, J. Elowitt and A. E. Clark, *WIREs Comput. Mol. Sci.*, 2024, **14**, e1729.
213 H. Wiener, *J. Am. Chem. Soc.*, 1947, **69**, 17–20.
214 A. A. Dobrynin, R. Entringer and I. Gutman, *Acta Appl. Math.*, 2001, **66**, 211–249.
215 H. Hosoya, *B. Chem. Soc. Jpn.*, 1971, **44**, 2332–2339.
216 H. Yang and W. Tian, in Proceedings of the 2024 8th International Conference on Electronic Information Technology and Computer Engineering, 2025.
217 M. Randic, *J. Am. Chem. Soc.*, 1975, **97**, 6609–6615.
218 Y. Shi, *Appl. Math. Comput.*, 2015, **265**, 1019–1025.
219 C. Dalfó, *Discrete. Math.*, 2019, **342**, 2792–2796.
220 G. Subashini, K. Kannan and A. Menaga, *Sci. Rep.*, 2024, **14**, 27214.
221 M. T. Farooq, N. Almalki and P. Kaemawichanurat, *Heliyon*, 2024, **10**, e37209.

222 S. Mondal, K. C. Das and D. Y. Huh, *Int. J. Quantum. Chem.*, 2024, **124**, e27336.
223 K. Aarthi, S. Elumalai, S. Balachandran and S. Mondal, *J. Appl. Math. Comput.*, 2025, **71**, 2727–2748.
224 A. G. Vrahatis, K. Lazaros and S. Kotsiantis, *Future Internet*, 2024, **16**, 318.
225 N. Yang, H. Wu, K. Zeng, Y. Li, S. Bao and J. Yan, *Fundam. Res.*, 2024, DOI: 10.1016/j.fmre.2024.11.027.
226 J.-N. Wu, T. Wang, Y. Chen, L.-J. Tang, H.-L. Wu and R.-Q. Yu, *Nat. Commun.*, 2024, **15**, 4993.
227 H. Guo, H. Zhu, G.-Y. Liu and Z.-X. Chen, *ACS Catal.*, 2024, **14**, 5720–5734.
228 M. Roucairol and T. Cazenave, *Mol. Inf.*, 2024, **43**, e202300259.
229 R. Mercado, E. J. Bjerrum and O. Engkvist, *J. Chem Inf. Model.*, 2022, **62**, 2093–2100.
230 X. Tang, J. Wang, M. Li, Y. He and Y. Pan, *Biomed. Res. Int.*, 2014, 354539.
231 Y. Zhao, M. Hayashida, J. Jindalertudomdee, H. Nagamochi and T. Akutsu, *J. Bioinform. Comput. Biol.*, 2013, **11**, 1343007.
232 C. Duran, G. Casadevall and S. Osuna, *Faraday. Discuss.*, 2024, **252**, 306–322.
233 Z. Zhang, S. Mohanty, J. Blanchet and W. Cai, *J. Mech. Phys. Solids.*, 2024, **187**, 105636.
234 A. W. M. Dress and T. F. Havel, *Discrete. Appl. Math.*, 1988, **19**, 129–144.
235 H.-C. Ehrlich and M. Rarey, *WIREs Comput. Mol. Sci.*, 2011, **1**, 68–79.
236 V. Bonnici, R. Giugno, A. Pulvirenti, D. Shasha and A. Ferro, *BMC Bioinformatics*, 2013, **14**, S13.
237 H.-C. Ehrlich and M. Rarey, *J. Cheminf.*, 2012, **4**, 13.
238 G. Corso, H. Stark, S. Jegelka, T. Jaakkola and R. Barzilay, *Nat. Rev. Method. Prim.*, 2024, **4**, 17.
239 F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner and G. Monfardini, *IEEE Trans. Neural Netw.*, 2009, **20**, 61–80.
240 V. Garg, *Curr. Opin. Struct. Biol.*, 2024, **84**, 102769.
241 M. Tang, B. Li and H. Chen, *Curr. Opin. Struct. Biol.*, 2023, **81**, 102616.
242 A. Mauri, V. Consonni and R. Todeschini, *Handbook of computational chemistry*, Springer International Publishing, Cham, 2017.
243 V. Consonni and R. Todeschini, *Molecular Descriptors*, Springer, Netherlands, Dordrecht, 2010.
244 T. Stepišnik, B. Škrlj, J. Wicker and D. Kocev, *Comput. Biol. Med.*, 2021, **130**, 104197.
245 Y. Zhao, R. J. Mulder, S. Houshyar and T. C. Le, *Polym. Chem.*, 2023, **14**, 3325–3346.
246 S. Kim, A. Jinich and A. Aspuru-Guzik, *J. Chem Inf. Model.*, 2017, **57**, 657–668.
247 A. Kubaib and P. M. Imran, *J. Mater. Sci.*, 2023, **58**, 4005–4019.
248 Z. Wang, L. Wang, H. Zhang, H. Xu and X. He, *Electron*, 2024, **2**, e41.
249 Z. Wang, L. Wang, H. Zhang, H. Xu and X. He, *Nano Converg.*, 2024, **11**, 8.
250 O. Allam, B. W. Cho, K. C. Kim and S. S. Jang, *RSC Adv.*, 2018, **8**, 39414–39420.
251 M. C. Potter, J. L. Goldberg and E. Aboufadel, *Advanced engineering mathematics*, Springer, Switzerland, 2005.
252 S. García, J. Luengo and F. Herrera, *Data preprocessing in data mining*, Springer, Switzerland, 2015.
253 H. Essén and M. Svensson, *Comput. Chem.*, 1996, **20**, 389–395.
254 H. B. Thompson, *J. Chem. Phys.*, 1967, **47**, 3407–3410.
255 V. Aquilanti and S. Cavalli, *J. Chem. Phys.*, 1986, **85**, 1355–1361.
256 D. J. Tobias and C. L. Brooks, III, *J. Chem. Phys.*, 1988, **89**, 5115–5127.
257 T. Engel, O. Sacher, A. Kolodzik, M. Rarey, J. A. de Sousa, T. Engel, C. Schwab and T. Engel, *Chemoinformatics. Basic Concepts and Methods*, Wiley-VCH, Weinheim, 2018.
258 A. Dalby, J. G. Nourse, W. D. Hounshell, A. K. I. Gushurst, D. L. Grier, B. A. Leland and J. Laufer, *J. Chem. Inf. Comp. Sci.*, 1992, **32**, 244–255.
259 G. S. Couch, E. F. Pettersen, C. C. Huang and T. E. Ferrin, *J. Mol. Graph.*, 1995, **13**, 153–158.
260 A. Horn and H. Lanig, *J. Mol. Model*, 1998, **5**, 141–142.
261 M. Rupp, A. Tkatchenko, K.-R. Müller and O. A. von Lilienfeld, *Phys. Rev. Lett.*, 2012, **108**, 058301.
262 C. R. Collins, G. J. Gordon, O. A. von Lilienfeld and D. J. Yaron, *J. Chem. Phys.*, 2018, **148**, 241718.
263 K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko and K.-R. Müller, *J. Chem. Theory Comput.*, 2013, **9**, 3404–3419.
264 O. Çaylak, O. Anatole von Lilienfeld and B. Baumeier, *Mach. Learn.: Sci. Technol.*, 2020, **1**, 03LT01.
265 J. Schrier, *J. Chem Inf. Model.*, 2020, **60**, 3804–3811.
266 F. Faber, A. Lindmaa, O. A. von Lilienfeld and R. Armiento, *Int. J. Quantum. Chem.*, 2015, **115**, 1094–1101.
267 K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller and A. Tkatchenko, *J. Phys. Chem. Lett.*, 2015, **6**, 2326–2331.
268 J. J. Sutherland, R. E. Higgs, I. Watson and M. Vieth, *J. Med. Chem.*, 2008, **51**, 2689–2700.
269 S. K. Chakravarti, *ACS Omega*, 2018, **3**, 2825–2836.
270 A. Varnek, D. Fourches, F. Hoonakker and V. P. Solov'ev, *J. Comput. Aid. Mol. Des.*, 2005, **19**, 693–703.
271 M. Congreve, G. Chessari, D. Tisi and A. J. Woodhead, *J. Med. Chem.*, 2008, **51**, 3661–3680.
272 A. H. Cheng, A. Cai, S. Miret, G. Malkomes, M. Phielipp and A. Aspuru-Guzik, *Digit. Discov.*, 2023, **2**, 748–758.
273 H. Öztürk, A. Özgür and E. Ozkirimli, *Bioinformatics*, 2018, **34**, i821–i829.
274 J. Degen, C. Wegscheid-Gerlach, A. Zaliani and M. Rarey, *ChemMedChem*, 2008, **3**, 1503–1507.
275 Y. Diao, F. Hu, Z. Shen and H. Li, *Bioinformatics*, 2023, **39**, btad012.
276 S. Jinsong, J. Qifeng, C. Xing, Y. Hao and L. Wang, *Commun. Chem.*, 2024, **7**, 20.
277 Q. Jia, Y. Zhang, Y. Wang, T. Ruan, M. Yao and L. Wang, *J. Mol. Graph. Model.*, 2025, **137**, 108985.
278 Z. Gao, X. Wang, B. Blumenfeld Gaines, X. Shi, J. Bi and M. Song, *Mol. Inf.*, 2023, **42**, 2200215.
279 B. Chen, Z. Pan, M. Mou, Y. Zhou and W. Fu, *Comput. Biol. Med.*, 2024, **169**, 107811.

280 M. Bon, A. Bilsland, J. Bower and K. McAulay, *Mol. Oncol.*, 2022, **16**, 3761–3777.
281 J. He, Y. Sun and J. Ling, *Interdiscip. Sci. Comput. Life Sci.*, 2025, **17**, 42–58.
282 M. Podda, D. Bacciu and A. Micheli, in International Conference on Artificial Intelligence and Statistics, 2020.
283 D. Flam-Shepherd, K. Zhu and A. Aspuru-Guzik, *Nat. Commun.*, 2022, **13**, 3293.
284 Z. Meng, C. Chen, X. Zhang, W. Zhao and X. Cui, *Big Data Min. Anal.*, 2024, **7**, 565–576.
285 J. Yoo, W. Jang and W.-H. Shin, *Curr. Opin. Struc. Biol.*, 2025, **91**, 102995.
286 I. Cortés-Ciriano and A. Bender, *J. Cheminf.*, 2019, **11**, 41.
287 M. Fernandez, F. Ban, G. Woo, M. Hsing, T. Yamazaki, E. LeBlanc, P. S. Rennie, W. J. Welch and A. Cherkasov, *J. Chem Inf. Model.*, 2018, **58**, 1533–1543.
288 X.-C. Zhang, J.-C. Yi, G.-P. Yang, C.-K. Wu, T.-J. Hou and D.-S. Cao, *Brief. Bioinform.*, 2022, **23**, bbac033.
289 I. Khokhlov, L. Krasnov, M. V. Fedorov and S. Sosnin, *Chem. Methods*, 2022, **2**, e202100069.
290 D.-A. Clevert, T. Le, R. Winter and F. Montanari, *Chem. Sci.*, 2021, **12**, 14174–14181.
291 J. Staker, K. Marshall, R. Abel and C. M. McQuaw, *J. Chem Inf. Model.*, 2019, **59**, 1017–1029.
292 K. Rajan, A. Zielesny and C. Steinbeck, *J. Cheminf.*, 2021, **13**, 61.
293 W. X. Shen, X. Zeng, F. Zhu, Y. L. Wang, C. Qin, Y. Tan, Y. Y. Jiang and Y. Z. Chen, *Nat. Mach. Intell.*, 2021, **3**, 334–343.
294 Y. Qian, X. Li, J. Wu, A. Zhou, Z. Xu and Q. Zhang, *J. Comput. Chem.*, 2022, **43**, 255–264.
295 Y. Chen, C. T. Leung, Y. Huang, J. Sun, H. Chen and H. Gao, *J. Cheminf.*, 2024, **16**, 141.
296 Y. Matsuzaka and Y. Uesawa, *Front. Bioeng. Biotechnol.*, 2019, **7**, 65.
297 S. Zhong, J. Hu, X. Yu and H. Zhang, *Chem. Eng. J.*, 2021, **408**, 127998.
298 S. Park and C. Seok, *J. Chem Inf. Model.*, 2022, **62**, 3157–3168.
299 T. Shi, Y. Yang, S. Huang, L. Chen, Z. Kuang, Y. Heng and H. Mei, *Chemom. Intell. Lab. Syst.*, 2019, **194**, 103853.
300 L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie and L. Farhan, *J. Big Data*, 2021, **8**, 53.
301 E. H. Herskovits, *Ann. Transl. Med.*, 2021, **9**, 824.
302 D. Kuzminykh, D. Polykovskiy, A. Kadurin, A. Zhebrak, I. Baskov, S. Nikolenko, R. Shayakhmetov and A. Zhavoronkov, *Mol. Pharm.*, 2018, **15**, 4378–4385.
303 C. Pang, H. H. Y. Tong and L. Wei, *Quant. Biol.*, 2023, **11**, 395–404.
304 A. Mauri and M. Bertola, *Int. J. Mol. Sci.*, 2022, **23**, 12882.
305 BlueDesc, **https://github.com/OlivierBeq/BlueDesc**, accessed 24 July 2025.
306 E. L. Willighagen, J. W. Mayfield, J. Alvarsson, A. Berg, L. Carlsson, N. Jeliazkova, S. Kuhn, T. Pluskal, M. Rojas-Chertó, O. Spjuth, G. Torrance, C. T. Evelo, R. Guha and C. Steinbeck, *J. Cheminf.*, 2017, **9**, 33.
307 J. Dong, D.-S. Cao, H.-Y. Miao, S. Liu, B.-C. Deng, Y.-H. Yun, N.-N. Wang, A.-P. Lu, W.-B. Zeng and A. F. Chen, *J. Cheminf.*, 2015, **7**, 60.
308 S. Höck and R. Riedl, *J. Cheminf.*, 2012, **4**, 38.
309 chemkit, **https://github.com/kylelutz/chemkit**, accessed 24 July 2025.
310 Y. Cao, A. Charisi, L.-C. Cheng, T. Jiang and T. Girke, *Bioinformatics*, 2008, **24**, 1733–1734.
311 D.-S. Cao, Q.-S. Xu, Q.-N. Hu and Y.-Z. Liang, *Bioinformatics*, 2013, **29**, 1092–1094.
312 J. Dong, Z.-J. Yao, M.-F. Zhu, N.-N. Wang, B. Lu, A. F. Chen, A.-P. Lu, H. Miao, W.-B. Zeng and D.-S. Cao, *J. Cheminf.*, 2017, **9**, 27.
313 F. Heidar-Zadeh, M. Richer, S. Fias, R. A. Miranda-Quintana, M. Chan, M. Franco-Pérez, C. E. González-Espinoza, T. D. Kim, C. Lanssens, A. H. G. Patel, X. D. Yang, E. Vöhringer-Martinez, C. Cárdenas, T. Verstraelen and P. W. Ayers, *Chem. Phys. Lett.*, 2016, **660**, 307–312.
314 N. M. O'Boyle and G. R. Hutchison, *Chem. Cent. J.*, 2008, **2**, 24.
315 Daylight, **https://www.daylight.com/products/index.html**, accessed 24 July 2025.
316 B. Ramsundar, *Deep Learning for the Life Sciences*, O'Reilly Media, 1st edn, 2019.
317 Dragon 7, **https://www.talete.mi.it/products/dragon_description.htm**, accessed 24 July 2025.
318 I. V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V. A. Palyulin, E. V. Radchenko, N. S. Zefirov, A. S. Makarenko, V. Y. Tanchuk and V. V. Prokopenko, *J. Comput. Aid. Mol. Des.*, 2005, **19**, 453–463.
319 G. Csardi and T. Nepusz, *Complex. Syst.*, 2006, **1695**, 1–9.
320 Indigo, **https://github.com/epam/Indigo**, accessed 24 July 2025.
321 G. Hinselmann, L. Rosenbaum, A. Jahn, N. Fechner and A. Zell, *J. Cheminf.*, 2011, **3**, 3.
322 H. Hong, Q. Xie, W. Ge, F. Qian, H. Fang, L. Shi, Z. Su, R. Perkins and W. Tong, *J. Chem Inf. Model.*, 2008, **48**, 1337–1344.
323 S. Doerr, M. J. Harvey, F. Noé and G. De Fabritiis, *J. Chem. Theory Comput.*, 2016, **12**, 1845–1852.
324 H. Moriwaki, Y.-S. Tian, N. Kawashita and T. Takagi, *J. Cheminf.*, 2018, **10**, 4.
325 A. A. Hagberg, D. A. Schult and P. Swart, in Proceedings of the 7th Python in Science Conference (SciPy2008), 2008.
326 N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, *J. Cheminf.*, 2011, **3**, 1–14.
327 J. Wahl and T. Sander, *J. Chem Inf. Model.*, 2022, **62**, 2202–2211.
328 OPSIN, **https://www.ebi.ac.uk/opsin**, accessed 24 July 2025.
329 C. W. Yap, *J. Comput. Chem.*, 2011, **32**, 1466–1474.
330 N. M. O'Boyle, C. Morley and G. R. Hutchison, *Chem. Cent. J.*, 2008, **2**, 5.
331 D.-S. Cao, Y.-Z. Liang, J. Yan, G.-S. Tan, Q.-S. Xu and S. Liu, *J. Chem Inf. Model.*, 2013, **53**, 3086–3096.
332 D.-S. Cao, N. Xiao, Q.-S. Xu and A. F. Chen, *Bioinformatics*, 2015, **31**, 279–281.

333 S. A. Rahman, M. Bashton, G. L. Holliday, R. Schrader and J. M. Thornton, *J. Cheminf.*, 2009, **1**, 12.
334 B. D. McKay, M. A. Yirik and C. Steinbeck, *J. Cheminf.*, 2022, **14**, 24.
335 A. Turing, *Mind*, 1950, **59**, 433–460.
336 J. McCarthy, M. L. Minsky, N. Rochester and C. E. Shannon, *Ai Mag.*, 2006, **27**, 12.
337 T. M. Mitchell, *Machine learning*, McGraw-Hill, New York, 1997.
338 Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, *Proc. IEEE*, 1998, **86**, 2278–2324.
339 G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
340 J. Redmon, S. Divvala, R. Girshick and A. Farhadi, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
341 P. M. Nadkarni, L. Ohno-Machado and W. W. Chapman, *J. Am. Med. Inform. Assn.*, 2011, **18**, 544–551.
342 L. Zhou, W. Schellaert, F. Martínez-Plumed, Y. Moros-Daval, C. Ferri and J. Hernández-Orallo, *Nature*, 2024, **634**, 61–68.
343 J. Hirschberg and C. D. Manning, *Science*, 2015, **349**, 261–266.
344 B. Horn, *Robot vision*, MIT press, Cambridge, 1986.
345 R. D. King, K. E. Whelan, F. M. Jones, P. G. K. Reiser, C. H. Bryant, S. H. Muggleton, D. B. Kell and S. G. Oliver, *Nature*, 2004, **427**, 247–252.
346 T. Brogårdh, *Annu. Rev. Control*, 2007, **31**, 69–79.
347 P. Viola and M. Jones, in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001.
348 C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
349 A. Voulodimos, N. Doulamis, A. Doulamis and E. Protopapadakis, *Comput. Intell. Neurosci.*, 2018, 7068349.
350 K. Chowdhary, *Natural language processing*, Springer, New Delhi, 2020.
351 S. Farquhar, J. Kossen, L. Kuhn and Y. Gal, *Nature*, 2024, **630**, 625–630.
352 L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, P.-A. Duquenne, H. Elsahar, H. Gong, K. Heffernan, J. Hoffman, C. Klaiber, P. Li, D. Licht, J. Maillard, A. Rakotoarison, K. R. Sadagopan, G. Wenzek, E. Ye, B. Akula, P.-J. Chen, N. El Hachem, B. Ellis, G. M. Gonzalez, J. Haaheim, P. Hansanti, R. Howes, B. Huang, M.-J. Hwang, H. Inaguma, S. Jain, E. Kalbassi, A. Kallet, I. Kulikov, J. Lam, D. Li, X. Ma, R. Mavlyutov, B. Peloquin, M. Ramadan, A. Ramakrishnan, A. Sun, K. Tran, T. Tran, I. Tufanov, V. Vogeti, C. Wood, Y. Yang, B. Yu, P. Andrews, C. Balioglu, M. R. Costa-jussà, O. Çelebi, M. Elbayad, C. Gao, F. Guzmán, J. Kao, A. Lee, A. Mourachko, J. Pino, S. Popuri, C. Ropers, S. Saleem, H. Schwenk, P. Tomasello, C. Wang, J. Wang, S. Wang and S. C. Team, *Nature*, 2025, **637**, 587–593.
353 J. M. Granda, L. Donina, V. Dragone, D.-L. Long and L. Cronin, *Nature*, 2018, **559**, 377–381.
354 A. Billard and D. Kragic, *Science*, 2019, **364**, eaat8414.
355 P. Slade, C. Atkeson, J. M. Donelan, H. Houdijk, K. A. Ingraham, M. Kim, K. Kong, K. L. Poggensee, R. Riener, M. Steinert, J. Zhang and S. H. Collins, *Nature*, 2024, **633**, 779–788.
356 D. E. Rumelhart, G. E. Hinton and R. J. Williams, *Nature*, 1986, **323**, 533–536.
357 A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, in *Neural Information Processing Systems*, 2017.
358 T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei, in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.
359 A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang and C. Ruan, *arXiv*, 2024, preprint, arXiv:2412.19437, DOI: **10.48550/arXiv.2412.19437**.
360 J. Li, K. Lim, H. Yang, Z. Ren, S. Raghavan, P.-Y. Chen, T. Buonassisi and X. Wang, *Matter*, 2020, **3**, 393–432.
361 B. L. DeCost, J. R. Hattrick-Simpers, Z. Trautt, A. G. Kusne, E. Campo and M. L. Green, *Mach. Learn.: Sci. Technol*, 2020, **1**, 033001.
362 K. Guo, Z. Yang, C.-H. Yu and M. J. Buehler, *Mater. Horizons*, 2021, **8**, 1153–1172.
363 E. O. Pyzer-Knapp, J. W. Pitera, P. W. J. Staar, S. Takeda, T. Laino, D. P. Sanders, J. Sexton, J. R. Smith and A. Curioni, *npj Comput. Mater.*, 2022, **8**, 84.
364 H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, S. Liu, P. Van Katwyk, A. Deac, A. Anandkumar, K. Bergen, C. P. Gomes, S. Ho, P. Kohli, J. Lasenby, J. Leskovec, T.-Y. Liu, A. Manrai, D. Marks, B. Ramsundar, L. Song, J. Sun, J. Tang, P. Veličković, M. Welling, L. Zhang, C. W. Coley, Y. Bengio and M. Zitnik, *Nature*, 2023, **620**, 47–60.
365 P. Schneider, W. P. Walters, A. T. Plowright, N. Sieroka, J. Listgarten, R. A. Goodnow, J. Fisher, J. M. Jansen, J. S. Duca, T. S. Rush, M. Zentgraf, J. E. Hill, E. Krutoholow, M. Kohler, J. Blaney, K. Funatsu, C. Luebkemann and G. Schneider, *Nat. Rev. Drug Discov.*, 2020, **19**, 353–364.
366 H. H. Loeffler, J. He, A. Tibo, J. P. Janet, A. Voronov, L. H. Mervin and O. Engkvist, *J. Cheminf.*, 2024, **16**, 20.
367 P. Xiao, X. Yun, Y. Chen, X. Guo, P. Gao, G. Zhou and C. Zheng, *Chem. Soc. Rev.*, 2023, **52**, 5255–5316.
368 Q. Meng, Y. Huang, L. Li, F. Wu and R. Chen, *Joule*, 2024, **8**, 344–373.
369 C. Lv, X. Zhou, L. Zhong, C. Yan, M. Srinivasan, Z. W. Seh, C. Liu, H. Pan, S. Li, Y. Wen and Q. Yan, *Adv. Mater.*, 2022, **34**, 2101474.
370 K. Liu, Z. Wei, C. Zhang, Y. Shang, R. Teodorescu and Q. L. Han, *IEEE/CAA J. Autom. Sinica*, 2022, **9**, 1139–1165.
371 S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen and B. Yu, *Nucleic. Acids. Res.*, 2023, **51**, D1373–D1380.

372 P. J. Linstrom and W. G. Mallard, *J. Chem. Eng. Data*, 2001, **46**, 1059–1063.
373 S. W. Gabrielson, *J. Med. Libr. Assoc.*, 2018, **106**, 588.
374 Y. Yang, K. U. Lao, D. M. Wilkins, A. Grisafi, M. Ceriotti and R. A. DiStasio, *Sci. Data*, 2019, **6**, 152.
375 J. Hoja, L. Medrano Sandonas, B. G. Ernst, A. Vazquez-Mayagoitia, R. A. DiStasio Jr and A. Tkatchenko, *Sci. Data*, 2021, **8**, 43.
376 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. von Lilienfeld, *Sci. Data*, 2014, **1**, 140022.
377 J. J. Valdés and A. B. Tchagang, *J. Comput. Chem.*, 2024, **45**, 1193–1214.
378 G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller and O. Anatole von Lilienfeld, *New J. Phys.*, 2013, **15**, 095003.
379 T. Fink, H. Bruggesser and J.-L. Reymond, *Angew. Chem., Int. Ed.*, 2005, **44**, 1504–1508.
380 T. Fink and J.-L. Reymond, *J. Chem Inf. Model.*, 2007, **47**, 342–353.
381 L. C. Blum and J.-L. Reymond, *J. Am. Chem. Soc.*, 2009, **131**, 8732–8733.
382 L. Ruddigkeit, R. van Deursen, L. C. Blum and J.-L. Reymond, *J. Chem Inf. Model.*, 2012, **52**, 2864–2875.
383 J. Hur and D. J. Wild, *Chem. Cent. J.*, 2008, **2**, 11.
384 M. R. Southern and P. R. Griffin, *Bioinformatics*, 2011, **27**, 741–742.
385 R. J. Bienstock, *Frontiers in Molecular Design and Chemical Information Science: Introduction*, American Chemical Society, Washington, DC, 2016.
386 Y. Kang, W. Lee, T. Bae, S. Han, H. Jang and J. Kim, *J. Am. Chem. Soc.*, 2025, **147**, 3943–3958.
387 R. Rodriguez-Esteban and M. Bundschus, *Drug Discov. Today*, 2016, **21**, 997–1002.
388 D. M. Jessop, S. E. Adams and P. Murray-Rust, *J. Cheminf.*, 2011, **3**, 40.
389 K. M. Hettne, A. J. Williams, E. M. van Mulligen, J. Kleinjans, V. Tkachenko and J. A. Kors, *J. Cheminf.*, 2010, **2**, 3.
390 Z. Zheng, O. Zhang, C. Borgs, J. T. Chayes and O. M. Yaghi, *J. Am. Chem. Soc.*, 2023, **145**, 18048–18062.
391 W. Zhang, Q. Wang, X. Kong, J. Xiong, S. Ni, D. Cao, B. Niu, M. Chen, Y. Li, R. Zhang, Y. Wang, L. Zhang, X. Li, Z. Xiong, Q. Shi, Z. Huang, Z. Fu and M. Zheng, *Chem. Sci.*, 2024, **15**, 10600–10611.
392 N. Schneider, N. Fechner, G. A. Landrum and N. Stiefl, *J. Chem Inf. Model.*, 2017, **57**, 1816–1831.
393 M. Krallinger, O. Rabal, A. Lourenço, J. Oyarzabal and A. Valencia, *Chem. Rev.*, 2017, **117**, 7673–7761.
394 J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
395 G. Xintong, W. Hongzhi, Y. Song and G. Hong, *Expert. Syst. Appl.*, 2014, **41**, 7987–7994.
396 A. Ghezzi, D. Gabelloni, A. Martini and A. Natalicchio, *Int. J. Manag. Rev.*, 2018, **20**, 343–363.
397 H. Garcia-Molina, M. Joglekar, A. Marcus, A. Parameswaran and V. Verroios, *IEEE Trans. Knowl. Data Eng.*, 2016, **28**, 901–911.
398 C. Chai, J. Fan, G. Li, J. Wang and Y. Zheng, in 2019 IEEE 35th International Conference on Data Engineering, 2019.
399 AISD HOMO–LUMO, **https://doi.org/10.13139/ORNLNCCS/1869409**, accessed 28 May 2025.
400 Battery electrolytes BatElyte, **https://ai2db.ai4ec.ac.cn/batelyte**, accessed 28 May 2025.
401 GDB-9-Ex: Quantum chemical prediction of UV/Vis absorption spectra for GDB-9 molecules, **https://www.osti.gov/dataexplorer/biblio/dataset/1890227**, accessed 28 May 2025.
402 Quantum-Machine, **https://quantum-machine.org/datasets**, accessed 28 May 2025.
403 The Materials Project, **https://next-gen.materialsproject.org/**, accessed 28 May 2025.
404 Ab initio accelerated. Accurate global machine learning force fields with hundreds of atoms, **https://www.sgdml.org/#datasets**, accessed 28 May 2025.
405 S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt and K.-R. Müller, *Sci. Adv.*, 2017, **3**, e1603015.
406 S. Chmiela, V. Vassilev-Galindo, O. T. Unke, A. Kabylda, H. E. Sauceda, A. Tkatchenko and K.-R. Müller, *Sci. Adv.*, 2023, **9**, eadf0873.
407 D. S. Levine, M. Shuaibi, E. W. C. Spotte-Smith, M. G. Taylor, M. R. Hasyim, K. Michel, I. Batatia, G. Csányi, M. Dzamba and P. Eastman, *arXiv*, 2025, preprint, arXiv:2505.08762, DOI: **10.48550/arXiv.2505.08762**.
408 ORNL_AISD-Ex: Quantum chemical prediction of UV/Vis absorption spectra for over 10 million organic molecules, **https://doi.ccs.ornl.gov/dataset/13423cfb-df80-541c-a3d9-a2f042fbe507**, accessed 28 May 2025.
409 M. Nakata, T. Shimazaki, M. Hashimoto and T. Maeda, *J. Chem Inf. Model.*, 2020, **60**, 5891–5899.
410 R. Ramakrishnan, M. Hartmann, E. Tapavicza and O. A. von Lilienfeld, *J. Chem. Phys.*, 2015, **143**, 084111.
411 AAT Bioquest, **https://www.aatbio.com/data-sets/boiling-point-bp-and-melting-point-mp-reference-table**, accessed 28 May 2025.
412 Release Announcement - ChemACX 24.24.4, **https://support.revvitysignals.com/hc/en-us/articles/32761356400532-Release-Announcement-ChemACX-24-24-4**, accessed 28 May 2025.
413 ChEMBL, **https://www.ebi.ac.uk/chembl**, accessed 28 May 2025.
414 ChemBridge, **https://chembridge.com**, accessed 28 May 2025.
415 Chemical Database, **https://www.chemdb.csdb.cn/chemdb/home**, accessed 28 May 2025.
416 Chemexper, **https://www.chemexper.com**, accessed 28 May 2025.
417 ChemSpider, **https://www.chemspider.com**, accessed 28 May 2025.
418 Compound Structure Database, **https://organchem.csdb.cn/scdb/main/str_introduce.asp**, accessed 28 May 2025.
419 The Merck Index Online, **https://merckindex.rsc.org**, accessed 28 May 2025.
420 NIST Chemistry WebBook, **https://webbook.nist.gov/chemistry**, accessed 28 May 2025.

421 Organic Compounds Database, **https://www.colby.edu/chemistry/cmp/cmp.html**, accessed 28 May 2025.
422 PubChem, **https://pubchem.ncbi.nlm.nih.gov/docs**, accessed 28 May 2025.
423 CAS SciFinder, **https://www.cas.org/solutions/cas-scifinder-discovery-platform/cas-scifinder**, accessed 28 May 2025.
424 B. I. Tingle, K. G. Tang, M. Castanon, J. J. Gutierrez, M. Khurelbaatar, C. Dandarchuluun, Y. S. Moroz and J. J. Irwin, *J. Chem Inf. Model.*, 2023, **63**, 1166–1176.
425 Chemical Data Explorer, **https://www.chemdx.org**, accessed 28 May 2025.
426 CHEMriya: Expanding Your Drug Discovery Horizons with 55 Billion Molecules, **https://www.otavachemicals.com/products/chemriya**, accessed 28 May 2025.
427 REAL Space, **https://enamine.net/compound-collections/real-compounds/real-space-navigator**, accessed 28 May 2025.
428 J. S. Delaney, *J. Chem. Inf. Comp. Sci.*, 2004, **44**, 1000–1005.
429 eXplore, **https://www.emolecules.com/explore**, accessed 28 May 2025.
430 D. L. Mobley and J. P. Guthrie, *J. Comput. Aid. Mol. Des.*, 2014, **28**, 711–720.
431 Freedom Space 3.0, **https://chem-space.com/compounds/freedom-space**, accessed 28 May 2025.
432 Internet Bond-energy Databank, **https://ibond.las.ac.cn**, accessed 28 May 2025.
433 Molecular Universe, **https://molecular-universe.ses.ai/map**, accessed 28 May 2025.
434 Ultimate 100+ million compounds, **https://ultimate.mcule.com**, accessed 28 May 2025.
435 Virtual Screening and Computational Drug Discovery Services, **https://wuxibiology.com/drug-discovery-services/hit-finding-and-screening-services/virtual-screening**, accessed 28 May 2025.
436 M. Shevlin, *ACS Med. Chem. Lett.*, 2017, **8**, 601–607.
437 N. S. Eyke, B. A. Koscher and K. F. Jensen, *Trends Chem.*, 2021, **3**, 120–132.
438 N. Vervoort, K. Goossens, M. Baeten and Q. Chen, *Anal. Sci. Adv.*, 2021, **2**, 109–127.
439 L. Su, M. Ferrandon, J. A. Kowalski, J. T. Vaughey and F. R. Brushett, *J. Electrochem. Soc.*, 2014, **161**, A1905.
440 O. Borodin, M. Olguin, C. E. Spear, K. W. Leiter and J. Knap, *Nanotechnology*, 2015, **26**, 354003.
441 M. A. Tudoran and M. V. Putz, *Curr. Org. Chem.*, 2015, **19**, 359–386.
442 C. Benecke, R. Grund, R. Hohberger, A. Kerber, R. Laue and T. Wieland, *Anal. Chim. Acta*, 1995, **314**, 141–147.
443 J. R. Boes, O. Mamun, K. Winther and T. Bligaard, *J. Phys. Chem. A*, 2019, **123**, 2281–2285.
444 V. Gillet, A. P. Johnson, P. Mata, S. Sike and P. Williams, *J. Comput. Aid. Mol. Des*, 1993, **7**, 127–153.
445 R. C. Glen and A. Payne, *J. Comput. Aid. Mol. Des*, 1995, **9**, 181–202.
446 C. Ji, Y. Zheng, R. Wang, Y. Cai and H. Wu, *IEEE Trans. Neural Netw. Learn. Syst.*, 2021, **34**, 2323–2337.
447 W. Jin, R. Barzilay and T. Jaakkola, in *International Conference on Machine Learning*, 2020.
448 H. Shang, Y. Tao, Y. Gao, C. Zhang and X. Wang, *IEEE Trans. Syst. Man Cybern.: Syst.*, 2014, **45**, 122–128.
449 G. R. Pastel, T. P. Pollard, O. Borodin and M. A. Schroeder, *Chem. Rev.*, 2025, **125**, 3059–3164.
450 A. Zheng and A. Casari, *Feature engineering for machine learning: principles and techniques for data scientists*, O'Reilly Media, Inc., Sebastopol, 2018.
451 N. Altman and M. Krzywinski, *Nat. Methods*, 2018, **15**, 399–400.
452 W. Liang, G. A. Tadesse, D. Ho, L. Fei-Fei, M. Zaharia, C. Zhang and J. Zou, *Nat. Mach. Intell.*, 2022, **4**, 669–677.
453 J. Benesty, J. Chen, Y. Huang and I. Cohen, *Pearson Correlation Coefficient*, Springer, Heidelberg, 2009.
454 A. Kraskov, H. Stögbauer and P. Grassberger, *Phys. Rev. E*, 2004, **69**, 066138.
455 A. E. Roth, *The Shapley Value: Essays in honor of Lloyd S. Shapley*, Cambridge University Press, Cambridge, 1988.
456 Q. Jia, H. Liu, X. Wang, Q. Tao, L. Zheng, J. Li, W. Wang, Z. Liu, X. Gu, T. Shen, S. Hou, Z. Jin and J. Ma, *Angew. Chem., Int. Ed.*, 2025, **64**, e202424493.
457 I. Guyon and A. Elisseeff, *J. Mach. Learn. Res.*, 2003, **3**, 1157–1182.
458 L. Sthle and S. Wold, *Chemom. Intell. Lab. Syst.*, 1989, **6**, 259–272.
459 C. Ding and H. Peng, *J. Bioinform. Comput. Biol.*, 2005, **3**, 185–205.
460 R. Tibshirani, *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, 1996, **58**, 267–288.
461 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, in *31st Annual Conference on Neural Information Processing Systems*, 2017.
462 M. B. Kursa and W. R. Rudnicki, *J. Stat. Softw.*, 2010, **36**, 1–13.
463 D. L. Donoho, *IEEE. T. Inform. Theory*, 2006, **52**, 1289–1306.
464 R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler and L. M. Ghiringhelli, *Phys. Rev. Mater.*, 2018, **2**, 083802.
465 P. H. Winston, *Artificial intelligence*, Addison-Wesley Longman Publishing Co., Inc., Boston, 1992.
466 E. B. Hunt, *Artificial intelligence*, Academic Press, Cambridge, 2014.
467 S. Das, A. Dey, A. Pal and N. Roy, *Int. J. Comput. Appl.*, 2015, **115**, 9.
468 A. Niculescu-Mizil and R. Caruana, in *Proceedings of the 22nd International Conference on Machine learning*, 2005.
469 R. Caruana and A. Niculescu-Mizil, in *Proceedings of the 23rd International Conference on Machine learning*, 2006.
470 P. Cunningham, M. Cord and S. J. Delany, *Supervised learning*, Springer, Switzerland, 2008.
471 J. G. Dy and C. E. Brodley, *J. Mach. Learn. Res.*, 2004, **5**, 845–889.
472 T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, R. Tibshirani and J. Friedman, *Unsupervised learning*, Springer, Switzerland, 2009.
473 S. Naeem, A. Ali, S. Anam and M. M. Ahmed, *Int. J. Com. Dig. Syst.*, 2023, **1**, 13.
474 L. P. Kaelbling, M. L. Littman and A. W. Moore, *J. Artif. Intell. Res.*, 1996, **4**, 237–285.

475 C. Szepesvári, *Algorithms for reinforcement learning*, Springer, Switzerland, 2022.
476 S. Milani, N. Topin, M. Veloso and F. Fang, *Acm Comput. Surv.*, 2024, **56**, 1–36.
477 M. F. A. Hady and F. Schwenker, *Semi-supervised learning*, Springer, Heidelberg, 2013.
478 J. E. Van Engelen and H. H. Hoos, *Mach. Learn.*, 2020, **109**, 373–440.
479 F. Daneshfar, S. Soleymanbaigi, P. Yamini and M. S. Amini, *Eng. Appl. Artif. Intel*, 2024, **133**, 108215.
480 A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee and F. Makedon, *Technologies*, 2020, **9**, 2.
481 J. Gui, T. Chen, J. Zhang, Q. Cao, Z. Sun, H. Luo and D. Tao, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024, **46**, 9052–9071.
482 A. Golbraikh and A. Tropsha, *Mol. Divers.*, 2000, **5**, 231–243.
483 A. Storkey, *Dataset shift in machine learning*, MIT Press, Cambridge, 2009.
484 A. Rácz, D. Bajusz and K. Héberger, *Molecules*, 2021, **26**, 1111.
485 J. T. Leonard and K. Roy, *QSAR Comb. Sci.*, 2006, **25**, 235–251.
486 P. Patil, P.-O. Bachant-Winner, B. Haibe-Kains and J. T. Leek, *Bioinformatics*, 2015, **31**, 2318–2323.
487 D. M. Hawkins, *J. Chem. Inf. Comp. Sci.*, 2004, **44**, 1–12.
488 X. Ying, *J. Phys.: Conf. Ser.*, 2019, **1168**, 022022.
489 Z. Liu, Z. Xu, J. Jin, Z. Shen and T. Darrell, in *International Conference on Machine Learning*, 2023.
490 W. M. Van der Aalst, V. Rubin, H. M. Verbeek, B. F. van Dongen, E. Kindler and C. W. Günther, *Softw. Syst. Model.*, 2010, **9**, 87–111.
491 G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever and R. R. Salakhutdinov, *arXiv*, 2012, preprint, arXiv:1207.0580, DOI: **10.48550/arXiv.1207.0580**.
492 L. Wan, M. Zeiler, S. Zhang, Y. Le Cun and R. Fergus, in *International Conference on Machine Learning*, 2013.
493 N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, *J. Mach. Learn. Res.*, 2014, **15**, 1929–1958.
494 D. Hendrycks, M. Mazeika, S. Kadavath and D. Song, in *Advances in Neural Information Processing Systems*, 2019.
495 J. R. Busemeyer and Y.-M. Wang, *J. Math. Psychol.*, 2000, **44**, 171–189.
496 L. P. Hansen and T. J. Sargent, *Robustness*, Princeton University Press, Princeton, 2008.
497 D. Hendrycks, K. Lee and M. Mazeika, in *International Conference on Machine Learning*, 2019.
498 A. Subbaswamy, R. Adams and S. Saria, in *International Conference on Artificial Intelligence and Statistics*, 2021.
499 C. Zhang, S. Bengio, M. Hardt, B. Recht and O. Vinyals, *Commun. ACM*, 2021, **64**, 107–115.
500 Y. Zhou, J. Shen and Y. Cheng, in *The Thirteenth International Conference on Learning Representations*, 2024.
501 Z. Yang, Y. Yu, C. You, J. Steinhardt and Y. Ma, in *International Conference on Machine Learning*, 2020.
502 W. Huang, Y. Shi, Z. Xiong and X. X. Zhu, in *European Conference on Computer Vision*, 2024.
503 E. Briscoe and J. Feldman, *Cognition*, 2011, **118**, 2–16.
504 S. Pal and S. K. Gauri, *Comput. Ind. Eng.*, 2010, **59**, 976–985.
505 A. Botchkarev, *arXiv*, 2018, preprint, arXiv:1809.03006, DOI: **10.48550/arXiv.1809.03006**.
506 A. V. Tatachar, *Int. Res. J. Eng. Technol.*, 2021, **8**, 2395.
507 Y.-G. Hsieh, G. Niu and M. Sugiyama, in *International Conference on Machine Learning*, 2019.
508 M. S. Pepe, T. Cai and G. Longton, *Biometrics*, 2006, **62**, 221–229.
509 D. J. Hand and C. Anagnostopoulos, *Pattern. Recogn. Lett.*, 2013, **34**, 492–495.
510 Y. Manzali, M. Chahhou and M. E. Mohajir, in *2017 International Conference on Wireless Technologies, Embedded and Intelligent Systems*, 2017.
511 R. Susmaga, in *Intelligent Information Processing and Web Mining*, 2004.
512 S. Visa, B. Ramsay, A. L. Ralescu and E. Van Der Knaap, *Maics*, 2011, **710**, 120–127.
513 L. Pereira and N. Nunes, in *2017 IEEE International Conference on Smart Grid Communications*, 2017.
514 M. Grandini, E. Bagli and G. Visani, *arXiv*, 2020, preprint, arXiv:2008.05756, DOI: **10.48550/arXiv.2008.05756**.
515 N. Usunier, D. Buffoni and P. Gallinari, in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009.
516 R. C. de Amorim and C. Hennig, *Inform. Sci.*, 2015, **324**, 126–145.
517 D. L. Davies and D. W. Bouldin, *IEEE Trans. Pattern Anal. Mach. Intell.*, 1979, **1**, 224–227.
518 L. Hubert and P. Arabie, *J. Classif.*, 1985, **2**, 193–218.
519 N. Veyrat-Charvillon and F.-X. Standaert, in *Cryptographic Hardware and Embedded Systems*, 2009.
520 J. A. Nelder and R. W. M. Wedderburn, *J. R. Stat. Soc. Ser. A-Stat. Soc.*, 1972, **135**, 370–384.
521 C. K. I. Williams and C. E. Rasmussen, in *Advances in Neural Information Processing*, 1996.
522 B. E. Boser, I. M. Guyon and V. N. Vapnik, in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 1992.
523 Y. B. Ian Goodfellow and A. Courville, *Deep Learning*, Beijing Shengtong Printing Co.,Ltd, Beijing, 2017.
524 C. Cortes and V. Vapnik, *Mach. Learn.*, 1995, **20**, 273–297.
525 A. J. Smola and B. Schölkopf, *Stat. Comput.*, 2004, **14**, 199–222.
526 L. Breiman, *Mach. Learn.*, 1996, **24**, 123–140.
527 H. Tin Kam, in Proceedings of the Third International Conference on Document Analysis and Recognition, 1995.
528 L. Breiman, *Mach. Learn.*, 2001, **45**, 5–32.
529 Y. Freund and R. E. Schapire, *J. Comput. Syst. Sci.*, 1997, **55**, 119–139.
530 T. Q. Chen, C. Guestrin and M. Assoc Comp, in 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
531 S. Badirli, X. Liu, Z. Xing, A. Bhowmik, K. Doan and S. S. Keerthi, *arXiv*, 2020, preprint, arXiv:2002.07971, DOI: **10.48550/arXiv.2002.07971**.
532 N. Altman and M. Krzywinski, *Nat. Methods*, 2018, **15**, 399–400.

533 H.-S. Park and C.-H. Jun, *Expert. Syst. Appl.*, 2009, **36**, 3336–3341.
534 R. Bro and A. K. Smilde, *Anal. Methods*, 2014, **6**, 2812–2831.
535 G. A. R. Hinton, in *Advances in Neural Information Processing Systems*, 2002.
536 L. van der Maaten and G. Hinton, *J. Mach. Learn. Res.*, 2008, **9**, 2579–2605.
537 J. Healy and L. McInnes, *Nat. Rev. Method. Prim.*, 2024, **4**, 82.
538 K. Hornik, M. Stinchcombe and H. White, *Neural Networks*, 1989, **2**, 359–366.
539 G. Cybenko, *Math. Control Signal Systems*, 1989, **2**, 303–314.
540 I. Wallach, M. Dzamba and A. Heifets, *arXiv*, 2015, preprint, arXiv:1510.02855, DOI: **10.48550/arXiv.1510.02855**.
541 G. B. Goh, C. Siegel, A. Vishnu, N. O. Hodas and N. Baker, *arXiv*, 2017, preprint, arXiv:1706.06689, DOI: **10.48550/arXiv.1706.06689**.
542 X. Zeng, H. Xiang, L. Yu, J. Wang, K. Li, R. Nussinov and F. Cheng, *Nat. Mach. Intell.*, 2022, **4**, 1004–1016.
543 J. C. Pereira, E. R. Caffarena and C. N. dos Santos, *J. Chem Inf. Model.*, 2016, **56**, 2495–2506.
544 S. Hochreiter and J. Schmidhuber, *Neural. Comput.*, 1997, **9**, 1735–1780.
545 F. A. Gers, J. Schmidhuber and F. Cummins, *Neural. Comput.*, 2000, **12**, 2451–2471.
546 M. Schuster and K. K. Paliwal, *IEEE. T. Signal Proces*, 1997, **45**, 2673–2681.
547 A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke and J. Schmidhuber, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2009, **31**, 855–868.
548 A. Graves, A. Mohamed and G. Hinton, in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.
549 F. Grisoni, M. Moret, R. Lingwood and G. Schneider, *J. Chem. Inf. Model.*, 2020, **60**, 1175–1183.
550 P.-C. Kotsias, J. Arús-Pous, H. Chen, O. Engkvist, C. Tyrchan and E. J. Bjerrum, *Nat. Mach. Intell.*, 2020, **2**, 254–265.
551 E. J. Bjerrum and R. Threlfall, *arXiv*, 2017, preprint, arXiv:1705.04612, DOI: **10.48550/arXiv.1705.04612**.
552 M. Popova, M. Shvets, J. Oliva and O. Isayev, *arXiv*, 2019, preprint, arXiv:1905.13372, DOI: **10.48550/arXiv.1905.13372**.
553 S. Zheng, X. Yan, Q. Gu, Y. Yang, Y. Du, Y. Lu and J. Xu, *J. Cheminf.*, 2019, **11**, 5.
554 J. Carracedo-Cosme, C. Romero-Muñiz, P. Pou and R. Pérez, *ACS Appl. Mater. Interfaces*, 2023, **15**, 22692–22704.
555 S. Ishida, T. Aasawat, M. Sumita, M. Katouda, T. Yoshizawa, K. Yoshizoe, K. Tsuda and K. Terayama, *WIREs Comput. Mol. Sci.*, 2023, **13**, e1680.
556 J. M. Stokes, K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, N. M. Donghia, C. R. MacNair, S. French, L. A. Carfrae, Z. Bloom-Ackermann, V. M. Tran, A. Chiappino-Pepe, A. H. Badran, I. W. Andrews, E. J. Chory, G. M. Church, E. D. Brown, T. S. Jaakkola, R. Barzilay and J. J. Collins, *Cell*, 2020, **180**, 688–702.
557 B. Perozzi, R. Al-Rfou and S. Skiena, in Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014.
558 A. Grover and J. Leskovec, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
559 T. N. Kipf and M. Welling, *arXiv*, 2016, preprint, arXiv:1609.02907, DOI: **10.48550/arXiv.1609.02907**.
560 W. Hamilton, Z. Ying and J. Leskovec, in *Advances in Neural Information Processing Systems*, 2017.
561 P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio and Y. Bengio, *arXiv*, 2017, preprint, arXiv:1710.10903, DOI: **10.48550/arXiv.1710.10903**.
562 J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals and G. E. Dahl, in *International Conference on Machine Learning*, 2017.
563 Y. Wang, Z. Li and A. Barati Farimani, *Graph Neural Networks for Molecules*, Springer International Publishing, Cham, 2023.
564 K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller and A. Tkatchenko, *Nat. Commun.*, 2017, **8**, 13890.
565 K. Schütt, P.-J. Kindermans, H. E. Sauceda Felix, S. Chmiela, A. Tkatchenko and K.-R. Müller, in *Advances in Neural Information Processing Systems*, 2017.
566 N. Lubbers, J. S. Smith and K. Barros, *J. Chem. Phys.*, 2018, **148**, 241715.
567 O. T. Unke and M. Meuwly, *J. Chem. Theory Comput.*, 2019, **15**, 3678–3693.
568 J. Gasteiger, J. Groß and S. Günnemann, *arXiv*, 2020, preprint, arXiv:2003.03123, DOI: **10.48550/arXiv.2003.03123**.
569 J. Gasteiger, S. Giri, J. T. Margraf and S. Günnemann, *arXiv*, 2020, preprint, arXiv:2011.14115, DOI: **10.48550/arXiv.2011.14115**.
570 Z. Qiao, M. Welborn, A. Anandkumar, F. R. Manby and T. F. Miller, III, *J. Chem. Phys.*, 2020, **153**, 124111.
571 Z. Qiao, A. S. Christensen, M. Welborn, F. R. Manby, A. Anandkumar and T. F. Miller, *Proc. Natl. Acad. Sci. U. S. A*, 2022, **119**, e2205221119.
572 Y. Liu, L. Wang, M. Liu, X. Zhang, B. Oztekin and S. Ji, *arXiv*, 2021, preprint, arXiv.2102.05013, DOI: **10.48550/arXiv.2102.05013**.
573 L. Wang, Y. Liu, Y. Lin, H. Liu and S. Ji, in *Advances in Neural Information Processing Systems*, 2022.
574 N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff and P. Riley, *arXiv*, 2018, preprint, arXiv.1802.08219, DOI: **10.48550/arXiv.1802.08219**.
575 K. Schütt, O. Unke and M. Gastegger, in *International Conference on Machine Learning*, 2021.
576 B. Anderson, T. S. Hy and R. Kondor, in *Advances in Neural Information Processing Systems*, 2019.
577 B. K. Miller, M. Geiger, T. E. Smidt and F. Noé, *arXiv*, 2020, preprint, arXiv:2008.08461, DOI: **10.48550/arXiv.2008.08461**.
578 V. G. Satorras, E. Hoogeboom and M. Welling, in *International Conference on Machine Learning*, 2021.
579 J. Gasteiger, F. Becker and S. Günnemann, in *Advances in Neural Information Processing Systems*, 2021.
580 W. Du, H. Zhang, Y. Du, Q. Meng, W. Chen, N. Zheng, B. Shao and T.-Y. Liu, in *International Conference on Machine Learning*, 2022.

581 I. Batatia, D. P. Kovacs, G. Simm, C. Ortner and G. Csányi, in *Advances in Neural Information Processing Systems*, 2022.
582 Y. Du, L. Wang, D. Feng, G. Wang, S. Ji, C. P. Gomes and Z.-M. Ma, in *Advances in Neural Information Processing Systems*, 2023.
583 Y. Liu, J. Cheng, H. Zhao, T. Xu, P. Zhao, F. Tsung, J. Li and Y. Rong, *arXiv*, 2023, preprint, arXiv:2308.13212, DOI: 10.48550/arXiv.2308.13212.
584 Z. Zheng, Y. Liu, J. Li, J. Yao and Y. Rong, in Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024.
585 P. N. Stuart and J. Russell, *Artificial Intelligence: A Modern Approach, Global Edition*, Pearson Education, Berkeley, 2021.
586 F. Fuchs, D. Worrall, V. Fischer and M. Welling, in *Advances in Neural Information Processing Systems*, 2020.
587 P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas and A. A. Lee, *ACS Cent. Sci.*, 2019, **5**, 1572–1583.
588 M. J. Hutchinson, C. Le Lan, S. Zaidi, E. Dupont, Y. W. Teh and H. Kim, in *International Conference on Machine Learning*, 2021.
589 F. Wu, Q. Zhang, D. Radev, J. Cui, W. Zhang, H. Xing, N. Zhang and H. Chen, *arXiv*, 2021, preprint, arXiv.2110.01191, DOI: 10.48550/arXiv.2110.01191.
590 S. Luo, T. Chen, Y. Xu, S. Zheng, T.-Y. Liu, L. Wang and D. He, *arXiv*, 2022, preprint, arXiv.2210.01765, DOI: 10.48550/arXiv.2210.01765.
591 C. Ying, T. Cai, S. Luo, S. Zheng, G. Ke, D. He, Y. Shen and T.-Y. Liu, in *Advances in Neural Information Processing Systems*, 2021.
592 Y. Shi, S. Zheng, G. Ke, Y. Shen, J. You, J. He, S. Luo, C. Liu, D. He and T.-Y. Liu, *arXiv*, 2022, preprint, arXiv:2203.04810, DOI: 10.48550/arXiv.2203.04810.
593 P. Thölke and G. De Fabritiis, *arXiv*, 2022, preprint, arXiv.2202.02541, DOI: 10.48550/arXiv.2202.02541.
594 R. P. Pelaez, G. Simeon, R. Galvelis, A. Mirarchi, P. Eastman, S. Doerr, P. Thölke, T. E. Markland and G. De Fabritiis, *J. Chem. Theory Comput.*, 2024, **20**, 4076–4087.
595 V. Bagal, R. Aggarwal, P. K. Vinod and U. D. Priyakumar, *J. Chem Inf. Model.*, 2022, **62**, 2064–2076.
596 S. Lu, Z. Gao, D. He, L. Zhang and G. Ke, *Nat. Commun.*, 2024, **15**, 7104.
597 X. Ji, Z. Wang, Z. Gao, H. Zheng, L. Zhang and G. Ke, *arXiv*, 2024, preprint, arXiv.2406.14969, DOI: 10.48550/arXiv.2406.14969.
598 S. Gong, Y. Zhang, Z. Mu, Z. Pu, H. Wang, X. Han, Z. Yu, M. Chen, T. Zheng, Z. Wang, L. Chen, Z. Yang, X. Wu, S. Shi, W. Gao, W. Yan and L. Xiang, *Nat. Mach. Intell.*, 2025, **7**, 543–552.
599 W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang and Z. Dong, *arXiv*, 2025, preprint, arXiv.2303.18223, DOI: 10.48550/arXiv.2303.18223.
600 J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu and D. Amodei, *arXiv*, 2020, preprint, arXiv.2001.08361, DOI: 10.48550/arXiv.2001.08361.
601 X. Bi, D. Chen, G. Chen, S. Chen, D. Dai, C. Deng, H. Ding, K. Dong, Q. Du and Z. Fu, *arXiv*, 2024, preprint, arXiv.2401.02954, DOI: 10.48550/arXiv.2401.02954.
602 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, in *Advances in neural information processing systems*, **30**, 2017.
603 J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.
604 Language models are unsupervised multitask learners, accessed 30 April, 2025.
605 J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai and Q. V. Le, *arXiv*, 2021, preprint, arXiv.2109.01652, DOI: 10.48550/arXiv.2109.01652.
606 J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le and D. Zhou, in *Advances in Neural Information Processing Systems*, **35**, 2022.
607 P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. T. Yih and T. Rocktäschel, in Advances in Neural Information Processing Systems, **33**, 2020.
608 Hello GPT-4o, https://openai.com/index/hello-gpt-4o, accessed 13 May 2024.
609 Claude 3.5 Sonnet, https://www.anthropic.com/news/claude-3-5-sonnet, accessed 21 June 2024.
610 Llama3, https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md, accessed 18 April 2024.
611 Llama3.1, https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md, accessed 23 July 2024.
612 Qwen2.5: A Party of Foundation Models!, https://qwenlm.github.io/blog/qwen2.5, accessed 19 September 2024.
613 Learning to reason with LLMs, https://openai.com/index/learning-to-reason-with-llms, accessed 12 September 2024.
614 A. Liu, B. Feng, B. Wang, B. Wang, B. Liu, C. Zhao, C. Dengr, C. Ruan, D. Dai and D. Guo, *arXiv*, 2024, preprint, arXiv:2405.04434, DOI: 10.48550/arXiv.2405.04434.
615 D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang and X. Bi, *arXiv*, 2025, preprint, arXiv:2501.12948, DOI: 10.48550/arXiv.2501.12948.
616 Y. Zheng, H. Y. Koh, J. Ju, A. T. N. Nguyen, L. T. May, G. I. Webb and S. Pan, *Nat. Mach. Intell.*, 2025, **7**, 437–447.
617 Y. Na, J. J. Kim, C. Park, J. Hwang, C. Kim, H. Lee and J. Lee, *Mater. Adv.*, 2025, **6**, 2543–2548.
618 S. Zhao, S. Chen, J. Zhou, C. Li, T. Tang, S. J. Harris, Y. Liu, J. Wan and X. Li, *Cell Rep. Phys. Sci.*, 2024, **5**, 101844.
619 R. Wang, M. Yang and Y. Shen, in *ICML 2024 Workshop on Efficient and Accessible Foundation Models for Biological Discovery*, 2024.
620 D. A. Boiko, R. MacKnight, B. Kline and G. Gomes, *Nature*, 2023, **624**, 570–578.
621 T. Song, M. Luo, X. Zhang, L. Chen, Y. Huang, J. Cao, Q. Zhu, D. Liu, B. Zhang, G. Zou, G. Zhang, F. Zhang, W. Shang, Y. Fu, J. Jiang and Y. Luo, *J. Am. Chem. Soc.*, 2025, **147**, 12534–12545.
622 L. Jia, É. Brémond, L. Zaida, B. Gaüzère, V. Tognetti and L. Joubert, *J. Comput. Chem.*, 2024, **45**, 2383–2396.

623 E. Hruska, A. Gale and F. Liu, *J. Chem. Theory Comput.*, 2022, **18**, 1096–1108.
624 F. Wang, Z. Ma and J. Cheng, *J. Am. Chem. Soc.*, 2024, **146**, 14566–14575.
625 F. Wang and J. Cheng, *J. Chem. Phys.*, 2022, **157**, 024103.
626 F. Wang and J. Cheng, *Chem. Sci.*, 2022, **13**, 11570–11576.
627 F. Wang and J. Cheng, *Chin. J. Struct. Chem.*, 2023, **42**, 100061.
628 F. Wang, Y. Sun and J. Cheng, *J. Am. Chem. Soc.*, 2023, **145**, 4056–4064.
629 P. Gao, D. Kochan, Y.-H. Tang, X. Yang and E. G. Saldanha, *J. Power Sources*, 2025, **629**, 236035.
630 O. Allam, R. Kuramshin, Z. Stoichev, B. W. Cho, S. W. Lee and S. S. Jang, *Mater. Today Energy*, 2020, **17**, 100482.
631 S. Xu, J. Liang, Y. Yu, R. Liu, Y. Xu, X. Zhu and Y. Zhao, *J. Phys. Chem. C*, 2021, **125**, 21352–21358.
632 W.-Z. Huang, P. Xu, X.-Y. Huang, C.-Z. Zhao, X. Bie, H. Zhang, A. Chen, E. Kuzmina, E. Karaseva, V. Kolosnitsyn, X. Zhai, T. Jiang, L.-Z. Fan, D. Wang and Q. Zhang, *MetalMat*, 2024, **1**, e6.
633 Y. Zhang and X. Xu, *Ind. Eng. Chem. Res.*, 2021, **60**, 343–354.
634 T.-T. Wu, G.-L. Dai, J.-J. Xu, F. Cao, X.-H. Zhang, Y. Zhao and Y.-M. Qian, *Rare Metals*, 2023, **42**, 3269–3303.
635 P. Peljo and H. H. Girault, *Energy Environ. Sci.*, 2018, **11**, 2306–2309.
636 M. Cocchi, P. G. De Benedetti, R. Seeber, L. Tassi and A. Ulrici, *J. Chem. Inf. Comp. Sci.*, 1999, **39**, 1190–1203.
637 R. C. Schweitzer and J. B. Morris, *Anal. Chim. Acta*, 1999, **384**, 285–303.
638 R. C. Schweitzer and J. B. Morris, *J. Chem. Inf. Comp. Sci.*, 2000, **40**, 1253–1261.
639 N. Yao, X. Chen, X. Shen, R. Zhang, Z.-H. Fu, X.-X. Ma, X.-Q. Zhang, B.-Q. Li and Q. Zhang, *Angew. Chem., Int. Ed.*, 2021, **60**, 21473–21478.
640 J. Liang, S. Xu, L. Hu, Y. Zhao and X. Zhu, *Mat. Chem. Front.*, 2021, **5**, 3823–3829.
641 H. Hu, Y. Shan, Q. Zhao, J. Wang, L. Wu and W. Liu, *J. Energy Chem.*, 2024, **98**, 374–382.
642 H. Luo, Q. Gou, Y. Zheng, K. Wang, R. Yuan, S. Zhang, W. Fang, Z. Luogu, Y. Hu, H. Mei, B. Song, K. Sun, J. Wang and M. Li, *ACS Nano*, 2025, **19**, 2427–2443.
643 Y.-F. Huang, T. Gu, G. Rui, P. Shi, W. Fu, L. Chen, X. Liu, J. Zeng, B. Kang, Z. Yan, F. J. Stadler, L. Zhu, F. Kang and Y.-B. He, *Energy Environ. Sci.*, 2021, **14**, 6021–6029.
644 E. Ascencio-Medina, S. He, A. Daghighi, K. Iduoku, G. M. Casanola-Martin, S. Arrasate, H. González-Díaz and B. Rasulev, *Polymers*, 2024, **16**, 2731.
645 L. Chen, C. Kim, R. Batra, J. P. Lightstone, C. Wu, Z. Li, A. A. Deshmukh, Y. Wang, H. D. Tran, P. Vashishta, G. A. Sotzing, Y. Cao and R. Ramprasad, *npj Comput. Mater.*, 2020, **6**, 61.
646 A. Mannodi-Kanakkithodi, G. Pilania, T. D. Huan, T. Lookman and R. Ramprasad, *Sci. Rep.*, 2016, **6**, 20952.
647 P. Zhou, Y. Xiang and K. Liu, *Energy Environ. Sci.*, 2024, **17**, 8057–8077.
648 S. Liu, X. Ji, N. Piao, J. Chen, N. Eidson, J. Xu, P. Wang, L. Chen, J. Zhang, T. Deng, S. Hou, T. Jin, H. Wan, J. Li, J. Tu and C. Wang, *Angew. Chem., Int. Ed.*, 2021, **60**, 3661–3671.
649 J. Chen, H. Zhang, M. Fang, C. Ke, S. Liu and J. Wang, *ACS Energy Lett.*, 2023, **8**, 1723–1734.
650 Z. Li, Y. Zhou, Y. Wang and Y.-C. Lu, *Adv. Energy Mater.*, 2019, **9**, 1802207.
651 M. He, L. Zhu, Y. Liu, Y. Jia, Y. Hao, G. Ye, X. Hong, Z. Xiao, Y. Ma, J. Chen, M. B. Shafqat and Q. Pang, *Angew. Chem., Int. Ed.*, 2025, **64**, e202415053.
652 M. Baek, H. Shin, K. Char and J. W. Choi, *Adv. Mater.*, 2020, **32**, 2005022.
653 L. Cao, D. Li, E. Hu, J. Xu, T. Deng, L. Ma, Y. Wang, X.-Q. Yang and C. Wang, *J. Am. Chem. Soc.*, 2020, **142**, 21404–21409.
654 F. Xu, W. Guo, F. Wang, L. Yao, H. Wang, F. Tang, Z. Gao, L. Zhang, W. E. Z.-Q. Tian and J. Cheng, *Nat. Comput. Sci.*, 2025, **5**, 292–300.
655 Q. You, Y. Sun, F. Wang, J. Cheng and F. Tang, *J. Am. Chem. Soc.*, 2025, **147**, 14667–14676.
656 N. Yao, X. Chen, S.-Y. Sun, Y.-C. Gao, L. Yu, Y.-B. Gao, W.-L. Li and Q. Zhang, *Chem*, 2025, **11**, 102254.
657 N. Yao, L. Yu, Z.-H. Fu, X. Shen, T.-Z. Hou, X. Liu, Y.-C. Gao, R. Zhang, C.-Z. Zhao, X. Chen and Q. Zhang, *Angew. Chem., Int. Ed.*, 2023, **62**, e202305331.
658 V. Goussard, F. Duprat, J.-L. Ploix, G. Dreyfus, V. Nardello-Rataj and J.-M. Aubry, *J. Chem Inf. Model.*, 2020, **60**, 2012–2023.
659 A. K. Chew, M. Sender, Z. Kaplan, A. Chandrasekaran, J. Chief Elk, A. R. Browning, H. S. Kwak, M. D. Halls and M. A. F. Afzal, *J. Cheminf.*, 2024, **16**, 31.
660 C. Bilodeau, A. Kazakov, S. Mukhopadhyay, J. Emerson, T. Kalantar, C. Muzny and K. Jensen, *Chem. Eng. J.*, 2023, **464**, 142454.
661 Z. Shi, F. Song, C. Ji, Y. Xiao, C. Peng and H. Liu, *Ind. Eng. Chem. Res.*, 2024, **63**, 4571–4584.
662 Z. Chen, J. Chen, Y. Qiu, J. Cheng, L. Chen, Z. Qi and Z. Song, *ACS Sustainable Chem. Eng.*, 2024, **12**, 6648–6658.
663 G. Bradford, J. Lopez, J. Ruza, M. A. Stolberg, R. Osterude, J. A. Johnson, R. Gomez-Bombarelli and Y. Shao-Horn, *ACS Cent. Sci.*, 2023, **9**, 206–216.
664 X. Kang, Z. Zhao, J. Qian and R. Muhammad Afzal, *Ind. Eng. Chem. Res.*, 2017, **56**, 11344–11351.
665 X. Ma, J. Yu, Y. Hu, J. Texter and F. Yan, *Ind. Chem. Mater.*, 2023, **1**, 39–59.
666 Y. Zhao, X. Zhang, L. Deng and S. Zhang, *Comput. Chem. Eng.*, 2016, **92**, 37–42.
667 B. Huwaimel, J. Alanazi, M. Alanazi, T. N. Alharby and F. Alshammari, *Sci. Rep.*, 2024, **14**, 31857.
668 P. Dhakal and J. K. Shah, *Mol. Syst. Des. Eng.*, 2022, **7**, 1344–1353.
669 L. Yang, S. I. Sandler, C. Peng, H. Liu and Y. Hu, *Ind. Eng. Chem. Res.*, 2010, **49**, 12596–12604.
670 A. U. Bhat, S. S. Merchant and S. S. Bhagwat, *Ind. Eng. Chem. Res.*, 2008, **47**, 920–925.
671 A. Kilic, O. Abdelaty, M. Zeeshan, A. Uzun, R. Yildirim and D. Eroglu, *Chem. Eng. J.*, 2024, **490**, 151562.

672 Z. Acar, P. Nguyen and K. C. Lau, *Appl. Sci.*, 2022, **12**, 2408.
673 Z. Dai, L. Wang, X. Lu and X. Ji, *Green Energy Environ.*, 2024, **9**, 1802–1811.
674 X. Liu, J. Yin, X. Zhang, W. Qiu, W. Jiang, M. Zhang, L. Zhu, H. Li and H. Li, *Chemistry*, 2024, **6**, 1552–1571.
675 L. D. Hughes, D. S. Palmer, F. Nigsch and J. B. O. Mitchell, *J. Chem Inf. Model.*, 2008, **48**, 220–232.
676 J. L. McDonagh, T. van Mourik and J. B. O. Mitchell, *Mol. Inf.*, 2015, **34**, 715–724.
677 T. Galeazzo and M. Shiraiwa, *Environ. Sci.: Atmos*, 2022, **2**, 362–374.
678 W. Mi, H. Chen, D. Zhu, T. Zhang and F. Qian, *Chem. Commun.*, 2021, **57**, 2633–2636.
679 V. Venkatraman, S. Evjen, H. K. Knuutila, A. Fiksdahl and B. K. Alsberg, *J. Mol. Liq.*, 2018, **264**, 318–326.
680 Q. Li, J. Ren, Y. Liu and Y. Zhou, *Int. J. Refrig.*, 2022, **143**, 28–36.
681 J. C. Dearden, *Environ. Toxicol. Chem.*, 2003, **22**, 1696–1709.
682 S. D. Groven, C. Desgranges and J. Delhommelle, *Fluid. Phase. Equilibr.*, 2019, **484**, 225–231.
683 L. M. Egolf, M. D. Wessel and P. C. Jurs, *J. Chem. Inf. Comp. Sci.*, 1994, **34**, 947–956.
684 G. Espinosa, D. Yaffe, Y. Cohen, A. Arenas and F. Giralt, *J. Chem. Inf. Comp. Sci.*, 2000, **40**, 859–879.
685 Y.-m Dai, Z.-p Zhu, Z. Cao, Y.-f Zhang, J.-l Zeng and X. Li, *J. Mol. Graph. Model.*, 2013, **44**, 113–119.
686 Y. Pan, J. Jiang and Z. Wang, *J. Hazard. Mater.*, 2007, **147**, 424–430.
687 X. Sun, N. J. Krakauer, A. Politowicz, W.-T. Chen, Q. Li, Z. Li, X. Shao, A. Sunaryo, M. Shen, J. Wang and D. Morgan, *Mol. Inf.*, 2020, **39**, 1900101.
688 Z. Wang, H. Wen, Y. Su, W. Shen, J. Ren, Y. Ma and J. Li, *Chem. Eng. Sci.*, 2022, **248**, 117219.
689 C. A. Bergström, U. Norinder, K. Luthman and P. Artursson, *J. Chem. Inf. Comp. Sci.*, 2003, **43**, 1177–1185.
690 M. Karthikeyan, R. C. Glen and A. Bender, *J. Chem Inf. Model.*, 2005, **45**, 581–590.
691 I. V. Tetko, Y. Sushko, S. Novotarskyi, L. Patiny, I. Kondratov, A. E. Petrenko, L. Charochkina and A. M. Asiri, *J. Chem Inf. Model.*, 2014, **54**, 3320–3329.
692 I. V. Tetko, D. M. Lowe and A. J. Williams, *J. Cheminf.*, 2016, **8**, 2.
693 D. E. Needham, I. C. Wei and P. G. Seybold, *J. Am. Chem. Soc.*, 1988, **110**, 4186–4194.
694 A. T. Balaban, N. Joshi, L. B. Kier and L. H. Hall, *J. Chem. Inf. Comp. Sci.*, 1992, **32**, 233–237.
695 A. R. Katritzky, L. Mu, V. S. Lobanov and M. Karelson, *J. Phys. Chem.*, 1996, **100**, 10400–10407.
696 A. R. Katritzky, V. S. Lobanov and M. Karelson, *J. Chem. Inf. Comp. Sci.*, 1998, **38**, 28–41.
697 F. Gharagheizi, S. A. Mirkhani, P. Ilani-Kashkouli, A. H. Mohammadi, D. Ramjugernath and D. Richon, *Fluid. Phase. Equilibr*, 2013, **354**, 250–258.
698 C. Qu, A. J. Kearsley, B. I. Schneider, W. Keyrouz and T. C. Allison, *J. Mol. Graph. Model.*, 2022, **112**, 108149.
699 A. R. Katritzky, R. Petrukhin, R. Jain and M. Karelson, *J. Chem. Inf. Comp. Sci.*, 2001, **41**, 1521–1530.
700 F. A. Carroll, C.-Y. Lin and F. H. Quina, *Energy Fuels*, 2010, **24**, 4854–4856.
701 F. A. Carroll, C.-Y. Lin and F. H. Quina, *Energy Fuels*, 2010, **24**, 392–395.
702 F. A. Carroll, C.-Y. Lin and F. H. Quina, *Ind. Eng. Chem. Res.*, 2011, **50**, 4796–4800.
703 A. R. Katritzky, I. B. Stoyanova-Slavova, D. A. Dobchev and M. Karelson, *J. Mol. Graph. Model.*, 2007, **26**, 529–536.
704 N. I. Zhokhova, I. I. Baskin, V. A. Palyulin, A. N. Zefirov and N. S. Zefirov, *Russ. Chem. Bull.*, 2003, **52**, 1885–1892.
705 F. Gharagheizi, R. F. Alamdari and M. T. Angaji, *Energy Fuels*, 2008, **22**, 1628–1635.
706 T. C. Le, M. Ballard, P. Casey, M. S. Liu and D. A. Winkler, *Mol. Inf.*, 2015, **34**, 18–27.
707 H. Li, J. Hao and S.-Z. Qiao, *Adv. Mater.*, 2024, **36**, 2411991.
708 T. Qin, H. Yang, L. Wang, W. Xue, N. Yao, Q. Li, X. Chen, X. Yang, X. Yu, Q. Zhang and H. Li, *Angew. Chem., Int. Ed.*, 2024, **63**, e202408902.
709 P. Kirkpatrick and C. Ellis, *Nature*, 2004, **432**, 823.
710 Q. Zhang, A. Khetan, E. Sorkun, F. Niu, A. Loss, I. Pucher and S. Er, *Energy Storage Mater.*, 2022, **47**, 167–177.
711 Y. Yang, N. Yao, Y.-C. Gao, X. Chen, Y.-X. Huang, S. Zhang, H.-B. Zhu, L. Xu, Y.-X. Yao, S.-J. Yang, Z. Liao, Z. Li, X.-F. Wen, P. Wu, T.-L. Song, J.-H. Yao, J.-K. Hu, C. Yan, J.-Q. Huang and Q. Zhang, *Angew. Chem., Int. Ed.*, 2025, **64**, e202505212.
712 J. Du, J. Guo, Q. Sun, W. Liu, T. Liu, G. Huang and X. Zhang, *J. Mater. Chem. A*, 2024, **12**, 12034–12042.
713 P. M. Tagade, S. P. Adiga, S. Pandian, M. S. Park, K. S. Hariharan and S. M. Kolake, *npj Comput. Mater.*, 2019, **5**, 127.
714 Z. Yang, W. Ye, X. Lei, D. Schweigert, H.-K. Kwon and A. Khajeh, *npj Comput. Mater.*, 2024, **10**, 296.
715 A. Khajeh, X. Lei, W. Ye, Z. Yang, L. Hung, D. Schweigert and H.-K. Kwon, *Digit. Discov.*, 2025, **4**, 11–20.
716 X. Chen, M. Liu, S. Yin, Y.-C. Gao, N. Yao and Q. Zhang, *Angew. Chem., Int. Ed.*, 2025, **64**, e202503105.
717 B. Burger, P. M. Maffettone, V. V. Gusev, C. M. Aitchison, Y. Bai, X. Wang, X. Li, B. M. Alston, B. Li, R. Clowes, N. Rankin, B. Harris, R. S. Sprick and A. I. Cooper, *Nature*, 2020, **583**, 237–241.
718 Q. Zhu, F. Zhang, Y. Huang, H. Xiao, L. Zhao, X. Zhang, T. Song, X. Tang, X. Li, G. He, B. Chong, J. Zhou, Y. Zhang, B. Zhang, J. Cao, M. Luo, S. Wang, G. Ye, W. Zhang, X. Chen, S. Cong, D. Zhou, H. Li, J. Li, G. Zou, W. Shang, J. Jiang and Y. Luo, *Natl. Sci. Rev*, 2022, **9**, nwac190.
719 Q. Zhu, Y. Huang, D. Zhou, L. Zhao, L. Guo, R. Yang, Z. Sun, M. Luo, F. Zhang, H. Xiao, X. Tang, X. Zhang, T. Song, X. Li, B. Chong, J. Zhou, Y. Zhang, B. Zhang, J. Cao, G. Zhang, S. Wang, G. Ye, W. Zhang, H. Zhao, S. Cong, H. Li, L.-L. Ling, Z. Zhang, W. Shang, J. Jiang and Y. Luo, *Nat. Synth.*, 2024, **3**, 319–328.
720 T. Dai, S. Vijayakrishnan, F. T. Szczypiński, J.-F. Ayme, E. Simaei, T. Fellowes, R. Clowes, L. Kotopanov, C. E. Shields, Z. Zhou, J. W. Ward and A. I. Cooper, *Nature*, 2024, **635**, 890–897.

721 B. A. Koscher, R. B. Canty, M. A. McDonald, K. P. Greenman, C. J. McGill, C. L. Bilodeau, W. Jin, H. Wu, F. H. Vermeire, B. Jin, T. Hart, T. Kulesza, S.-C. Li, T. S. Jaakkola, R. Barzilay, R. Gómez-Bombarelli, W. H. Green and K. F. Jensen, *Science*, 2023, **382**, eadi1407.
722 N. J. Szymanski, B. Rendy, Y. Fei, R. E. Kumar, T. He, D. Milsted, M. J. McDermott, M. Gallant, E. D. Cubuk, A. Merchant, H. Kim, A. Jain, C. J. Bartel, K. Persson, Y. Zeng and G. Ceder, *Nature*, 2023, **624**, 86–91.
723 G. Tom, S. P. Schmid, S. G. Baird, Y. Cao, K. Darvish, H. Hao, S. Lo, S. Pablo-García, E. M. Rajaonson, M. Skreta, N. Yoshikawa, S. Corapi, G. D. Akkoc, F. Strieth-Kalthoff, M. Seifrid and A. Aspuru-Guzik, *Chem. Rev.*, 2024, **124**, 9633–9732.
724 A. Narayanan Krishnamoorthy, C. Wölke, D. Diddens, M. Maiti, Y. Mabrouk, P. Yan, M. Grünebaum, M. Winter, A. Heuer and I. Cekic-Laskovic, *Chem. Methods*, 2022, **2**, e202200008.
725 P. Yan, M. Fischer, H. Martin, C. Wölke, A. N. Krishnamoorthy, I. Cekic-Laskovic, D. Diddens, M. Winter and A. Heuer, *J. Mater. Chem. A*, 2024, **12**, 19123–19136.
726 J. Noh, H. A. Doan, H. Job, L. A. Robertson, L. Zhang, R. S. Assary, K. Mueller, V. Murugesan and Y. Liang, *Nat. Commun.*, 2024, **15**, 2757.
727 M. A. Stolberg, J. Lopez, S. D. Cawthern, A. Herzog-Arbeitman, H.-K. Kwon, D. Schweigert, A. Anapolosky, B. D. Storey, J. A. Johnson and Y. Shao-Horn, *Matter*, 2025, **8**, 102129.
728 M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao and B. Mons, *Sci. Data*, 2016, **3**, 160018.
729 B. Yurdem, M. Kuzlu, M. K. Gullu, F. O. Catak and M. Tabassum, *Heliyon*, 2024, **10**, e38137.
730 Y. Xu, X. Liu, W. Xia, J. Ge, C.-W. Ju, H. Zhang and J. Z. H. Zhang, *J. Chem Inf. Model.*, 2024, **64**, 8440–8452.
731 Q. Lv, G. Chen, Z. Yang, W. Zhong and C. Y. C. Chen, *IEEE Trans. Neural Netw. Learn. Syst.*, 2024, **35**, 11218–11230.
732 H. Zhao, S. Liu, M. Chang, H. Xu, J. Fu, Z. Deng, L. Kong and Q. Liu, in *Advances in neural information processing systems*, 2023.
733 J. Park, F. Sorourifar, M. R. Muthyala, A. M. Houser, M. Tuttle, J. A. Paulson and S. Zhang, *J. Am. Chem. Soc.*, 2024, **146**, 31230–31239.
734 Y. Wang, Z. Yang and Q. Yao, *Commun. Med.*, 2024, **4**, 59.
735 Y. Zhang, Q. Yao, L. Yue, X. Wu, Z. Zhang, Z. Lin and Y. Zheng, *Nat. Comput. Sci.*, 2023, **3**, 1023–1033.
736 Z. Wang, M. Chen, J. Wu, X. Ji, L. Zeng, J. Peng, J. Yan, A. A. Kornyshev, B. Mao and G. Feng, *Phys. Rev. Lett.*, 2025, **134**, 046201.
737 L. Zeng, M. Chen, Z. Wang, R. Qiao and G. Feng, *Phys. Rev. Lett.*, 2023, **131**, 096201.
738 S. Bi, H. Banda, M. Chen, L. Niu, M. Chen, T. Wu, J. Wang, R. Wang, J. Feng, T. Chen, M. Dincă, A. A. Kornyshev and G. Feng, *Nat. Mater.*, 2020, **19**, 552–558.
739 S. Kondrat, G. Feng, F. Bresme, M. Urbakh and A. A. Kornyshev, *Chem. Rev.*, 2023, **123**, 6668–6715.
740 X. Yu, M. Chen, Z. Li, X. Tan, H. Zhang, J. Wang, Y. Tang, J. Xu, W. Yin, Y. Yang, D. Chao, F. Wang, Y. Zou, G. Feng, Y. Qiao, H. Zhou and S.-G. Sun, *J. Am. Chem. Soc.*, 2024, **146**, 17103–17113.
741 Y. Liu, P. Yu, Q. Sun, Y. Wu, M. Xie, H. Yang, T. Cheng and W. A. Goddard, III, *ACS Energy Lett.*, 2021, **6**, 2320–2327.
742 Y. Chen, M. Li, Y. Liu, Y. Jie, W. Li, F. Huang, X. Li, Z. He, X. Ren, Y. Chen, X. Meng, T. Cheng, M. Gu, S. Jiao and R. Cao, *Nat. Commun.*, 2023, **14**, 2655.
743 Y. Chen, Y. Liu, Z. He, L. Xu, P. Yu, Q. Sun, W. Li, Y. Jie, R. Cao and T. Cheng, *Nat. Sci. Open.*, 2024, **3**, 20230039.
744 M. Company, Global Energy Perspective 2024, McKinsey & Company, 2024.