SCIENCE CHINA
Chemistry

•INVITED REVIEWS•

August 2024 Vol.67 No.8: 2461–2496
https://doi.org/10.1007/s11426-024-2072-4

AI for organic and polymer synthesis

Xin Hong[1*], Qi Yang[2*], Kuangbiao Liao[3*], Jianfeng Pei[4*], Mao Chen[5*], Fanyang Mo[6,8*], Hua Lu[7*], Wen-Bin Zhang[7,8*], Haisen Zhou[7], Jiaxiao Chen[4], Lebin Su[3], Shuo-Qing Zhang[1], Siyuan Liu[2], Xu Huang[9], Yi-Zhou Sun[1], Yuxiang Wang[7,8], Zexi Zhang[5], Zhunzhun Yu[3], Sanzhong Luo[2*], Xue-Feng Fu[10*] & Shu-Li You[11*]

[1]*Center of Chemistry for Frontier Technologies, Department of Chemistry, Zhejiang University, Hangzhou 310027, China;*
[2]*Center of Basic Molecular Science, Department of Chemistry, Tsinghua University, Beijing 100084, China;*
[3]*Guangzhou National Laboratory, Guangzhou 510005, China;*
[4]*Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China;*
[5]*State Key Laboratory of Molecular Engineering of Polymers, Department of Macromolecular Science, Fudan University, Shanghai 200433, China;*
[6]*School of Materials Science and Engineering, Peking University, Beijing 100871, China;*
[7]*Beijing National Laboratory for Molecular Sciences, Center for Soft Matter Science and Engineering, Key Laboratory of Polymer Chemistry and Physics of Ministry of Education, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China;*
[8]*AI for Science (AI4S)-Preferred Program, Shenzhen Graduate School, Peking University, Shenzhen 518055, China;*
[9]*State Key Laboratory of Chemical Biology, Shanghai Institute of Organic Chemistry, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200032, China;*
[10]*Department of Chemical Sciences, National Natural Science Foundation of China, Beijing 100085, China;*
[11]*State Key Laboratory of Organometallic Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai 200032, China*

Received March 20, 2024; accepted April 28, 2024; published online June 26, 2024

Recent years have witnessed the transformative impact from the integration of artificial intelligence with organic and polymer synthesis. This synergy offers innovative and intelligent solutions to a range of classic problems in synthetic chemistry. These exciting advancements include the prediction of molecular property, multi-step retrosynthetic pathway planning, elucidation of the structure-performance relationship of single-step transformation, establishment of the quantitative linkage between polymer structures and their functions, design and optimization of polymerization process, prediction of the structure and sequence of biological macromolecules, as well as automated and intelligent synthesis platforms. Chemists can now explore synthetic chemistry with unprecedented precision and efficiency, creating novel reactions, catalysts, and polymer materials under the data-driven paradigm. Despite these thrilling developments, the field of artificial intelligence (AI) synthetic chemistry is still in its infancy, facing challenges and limitations in terms of data openness, model interpretability, as well as software and hardware support. This review aims to provide an overview of the current progress, key challenges, and future development suggestions in the interdisciplinary field between AI and synthetic chemistry. It is hoped that this overview will offer readers a comprehensive understanding of this emerging field, inspiring and promoting further scientific research and development.

organic synthesis, polymer synthesis, machine learning prediction, chemical database, automated synthesis

Citation: Hong X, Yang Q, Liao K, Pei J, Chen M, Mo F, Lu H, Zhang WB, Zhou H, Chen J, Su L, Zhang SQ, Liu S, Huang X, Sun YZ, Wang Y, Zhang Z, Yu Z, Luo S, Fu XF, You SL. AI for organic and polymer synthesis. *Sci China Chem*, 2024, 67: 2461–2496, https://doi.org/10.1007/s11426-024-2072-4

*Corresponding authors (email: hxchem@zju.edu.cn; yang.q.17@outlook.com; kuangbiao@gzlab.ac.cn; jfpei@pku.edu.cn; chenmao@fudan.edu.cn; fmo@pku.edu.cn; chemhualu@pku.edu.cn; wenbin@pku.edu.cn; luosz@tsinghua.edu.cn; fuxf@nsfc.gov.cn; slyou@sioc.ac.cn)

CONTENTS

1 Introduction

Artificial intelligence (AI) encompasses a broad set of technologies that simulate human intelligence, of which machine learning (ML) is a crucial subset. ML enables computer systems to learn from and interpret data without explicit programming, forming the core mechanism behind most AI applications. By observing and analyzing massive datasets, AI algorithms can identify patterns, classify information, and even make complex decisions. Particularly in the field of natural language processing (NLP), the development of AI has been transformative. Recent years have seen large language models (LLMs) [1–4], like ChatGPT [5], significantly contribute to this advancement, embodying the dream of artificial general intelligence. In the field of chemistry, LLMs also have exciting applications: research has shown that LLMs inherently possess a certain degree of chemical understanding [6]. Models like Coscientist, which can autonomously design, plan, and execute chemical research, illustrate how LLMs facilitate chemical research by automating literature analysis and experimental processes [7]. This AI wave continues to expand, driving technological advancements across a broad spectrum of domains; for example, the advent of AlphaGo has seen it defeat top human players in Go [8], and the emergence of AlphaFold [9] signals the inevitable embrace of the AI revolution in natural sciences. This paradigm shift brought about by AI is profoundly influencing the methods through which humanity tackles and resolves complex high-dimensional problems.

Within the realm of synthetic chemistry, chemists are constantly faced with the challenges of complexity and multidimensionality. The inherent intricacy of these problems renders bottom-up theoretical deductions difficult, leading synthetic chemists to realize the potential of approaching these issues from the perspectives of data science and information science [10]. Whether in synthetic pathway planning [11–13] or exploring substituent effects [14], chemists have already widely applied data-driven methods. These methods, ranging from simple linear fitting to the development of complex expert systems, offered a powerful strategy for chemists to find solutions in the ocean of data, yielding fruitful advances in synthetic chemistry.

From the research journey of the substituent effect, we can appreciate the profound impact of data and intelligence on synthetic chemistry. The pioneering explorations of Ingold [15] and Robinson *et al*. [16] laid the foundation for concepts such as steric hindrance and electronic effects, now fundamental in organic chemistry textbooks. Hammett's systematic and in-depth application of linear relationship to the study of substituent effects has made the Hammett equation a cornerstone for analyzing organic reaction mechanisms [17,18]. Later, the exciting advances from Sigman and others revealed the potential of multivariate linear free energy relationship (LFER) in propelling the understanding and design of modern synthetic transformations [19]. Today, the continuous expansion of synthetic chemistry databases and the advancement of AI algorithms have enabled chemists to make chemical accuracy-level predictions of molecular properties directly from topological structures. A notable example includes the application of the *i*BonD database [20] for pK_a predictions, which matches the accuracy of quantum calculations while significantly enhancing efficiency by orders of magnitude [21]. These advancements have not only facilitated progress in organic synthesis, but also led to significant achievements in polymer synthesis and automated experimentation with important directions highlighted in Figure 1, signaling the dawn of a new era of intelligent synthesis. In this review, we focus not only on representative directions and outcomes of the intersection between artificial intelligence and synthetic chemistry, but also delve into the current challenges facing the field along with potential solutions. This list is by no means comprehensive, but it is our hope that through this review, readers will gain a clear and comprehensive perspective on the breadth and depth of AI applications in synthetic chemistry.

2 Machine learning pipeline

Prior to delving into specific research advancements, it is essential to elucidate the fundamentals of ML, especially as

they pertain to applications within synthetic chemistry. ML techniques are categorized into three primary types: supervised learning, where the goal is to learn a function mapping inputs to outputs given labeled data; unsupervised learning, aimed at uncovering the hidden structure of unlabeled data; and reinforcement learning, focused on learning how to take actions to maximize some notion of cumulative reward through interaction with an environment. ML typically encompasses four critical stages: data collection, encoding, model training, and result analysis (Figure 2). Initially, the collection and organization of relevant data lay the groundwork for model construction. Subsequently, during the encoding phase, these data are transformed into a format interpretable by ML models. The model training stage then utilizes encoded data, allowing algorithms to identify patterns and relationships within the data. Finally, result analysis evaluates the predictive performance of the model as well as interprets the rationale behind the model predictions. These stages collectively form the foundation for applying ML in the realm of synthetic chemistry, promoting the intelligent solution of chemical problems.

The primary sources of data in synthetic chemistry currently include public databases, high-throughput experimentation (HTE), computational simulations, and electronic laboratory notebooks (ELNs). These diverse data streams are vital for the success of ML modeling, offering extensive information on reactions, compounds, and properties. Table 1 lists exemplary open-access databases for organic and polymer synthesis. Public databases like Reaxys and Scifinder are indispensable for providing comprehensive chemical data, while HTE systems enable the efficient generation of large datasets through automated experiments, assessing thousands of reactions with minimal material and time. ELNs play a crucial role in documenting and sharing experimental details [22,23], although they present challenges related to data standardization, privacy, and variability. Together, these sources underpin the development of ML models in synthetic chemistry, leveraging the vast array of data to fuel innovations through computational analysis and experimental integration.

Encoding molecules and reactions into machine-readable formats are critical for ML modeling. The molecular encodings can be characterized by a hierarchy of complexity: zero-dimensional physicochemical properties like molecular

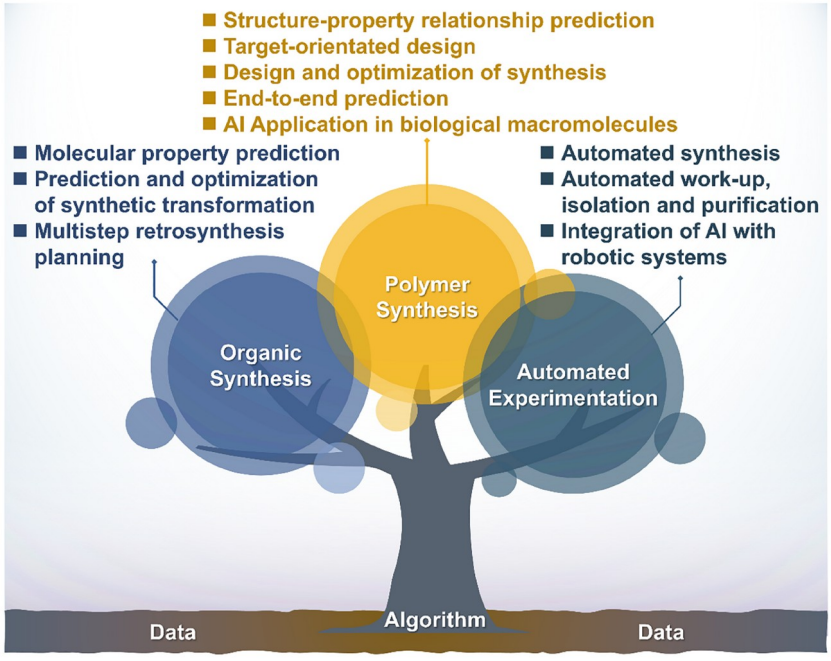

Figure 1 Representative research directions of AI applications in organic and polymer synthesis (color online).

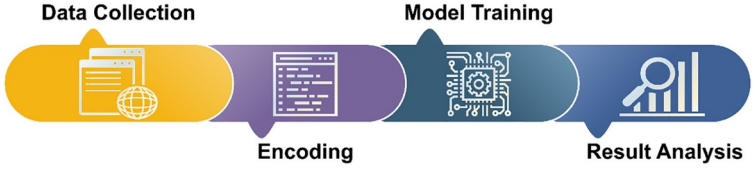

Figure 2 Typical pipeline of machine learning modeling (color online).

Table 1 Overview of representative open access databases for organic and polymer synthesis

Database	Description	Database	Description
ChEBI [24]	Molecular entities of small chemical compounds.	*I*BonD [20]	Chemical database that covers heterolytic ($pK_{a)}$ and homolytic bond dissociation energies (BDE).
ChEMBL [25]	Database of bioactive molecules with drug-like properties.	SDBS [26]	Spectra database system for organic compounds.
COD [27]	Crystal structure database of organic, inorganic, and metal-organic compounds.	SpectraBase [28]	Spectra database for organic, organometallic, and inorganic compounds.
NIST chemistry Webbook [29]	NIST standard reference database of chemical and physical property data.	UniChem [30]	Database of pointers between chemical structures and EMBL-EBI chemistry resources.
OSCAR [31]	Datasets of chemically and functionally diverse organocatalysts.	ZINC20 [32]	Database of commercially available compounds.
PubChem [33,34]	Collection of freely accessible chemical information.	ORD [35]	Open organic reaction database.
ChemSpider	Database of chemical information including molecular structures and properties.	PoLyInfo	Polymer database that covers properties, structures, processing methods, *etc*.
USPTO	Open data of United States patents.	MatWeb	Material property database that includes information on a wide range of materials and polymers.
Quantum-machine [36,37]	Quantum chemistry calculation database.	Synthesis Explorer	Curated collection of chemical reactions and synthesis pathways.

weight and Log*P*, one-dimensional string representations such as SMILES [38] and SELFIES [39] for encoding atomic types and connections, two-dimensional molecular fingerprints capturing molecular structures without stereochemical details, and three-dimensional descriptors that include stereochemistry and quantum chemical features for a comprehensive representation of molecular conformations. Additionally, graph-based learning methods offer advanced ways to depict molecules and reactions [40], addressing the complexity of chemical structures in multidimensional space. For polymers, the challenge of their stochastic nature is met with novel encoding strategies like BigSMILES [41] for sequence distributions and PolyGrammar [42] for hypergraph representations, alongside graph neural networks and Transformer-based language models to distinguish polymer sequences and topologies. These encoding strategies are essential for effectively processing and analyzing chemical data, enabling the advancement of ML applications in organic and polymer systems.

The process of ML modeling involves utilizing algorithms to grasp patterns within data, thereby enabling predictions about unknown targets. This involves a spectrum of methodologies from classical ML, adept at handling linear relationships and structured data, to deep learning, known for its proficiency with large-scale and complex datasets. Classical models like linear, tree-based, and kernel-based methods offer solutions for simpler relationships, while deep learning's layered architecture allows for the extraction of high-dimensional features. In addition, the modern ML process can be adaptive with active learning and transfer learning techniques. Active learning dynamically selects the most informative data points for labeling and training, effectively improving model performance with less data. Transfer learning leverages knowledge acquired from one domain to enhance model accuracy in another, significantly reducing the need for extensive labeled datasets in new applications. Selecting the right model and algorithm is crucial, depending on the data's nature and the analytical task at hand. Typical evaluation methods include cross-validation and independent test sets, which are essential to ensure the model's generalizability and effectiveness in real-world applications.

For ML applications, result analysis and model interpretation are equally important [43,44], as merely relying on and executing each model prediction is insufficient and lacks comprehensiveness. In the analysis of ML predictions, one can leverage domain expertise to evaluate and contrast predictions against non-ML methodologies. Techniques such as sensitivity analysis and hypothesis testing further aid in assessing the accuracy and reliability of models, especially when predictions deviate from expected norms. Additionally, methods like dimensionality reduction and clustering are instrumental in deriving valuable insights from ML predictions, as demonstrated in Tim Cernak's utilization of graph editing distance to analyze retrosynthesis routes designed by SYNTHIA [45]. The goal of model interpretation is to elucidate the decision-making process of complex, high-dimensional ML models, thereby extracting heuristic principles and knowledge pertinent to the domain. Crucial approaches include feature importance analysis, which pinpoints key influencing variables, and interpretability frameworks such as SHAP [46,47] and LIME [48,49], which facilitate the understanding of how models arrive at their decisions.

3 AI applications in organic synthesis

In the realm of organic synthesis, data-driven methodologies are catalyzing solutions to a multitude of complex challenges, achieving substantial progress in recent years. These issues encompass a vast array of dimensions, from single molecule to intermolecular chemical reactions, and even multi-step transformations in total synthesis. The scope of research in this area is equally comprehensive, involving the prediction of molecular physicochemical properties, the evaluation of structure-activity relationships in organic transformations, and the optimization of reaction conditions. Facing these chemical challenges across diverse scenarios, researchers have applied and developed innovative AI technologies, with key directions highlighted in Figure 3. These successes demonstrate the immense potential of AI within the sphere of organic synthesis. In this section, the representative research advances are discussed to showcase the ability of AI technology in these application scenarios.

3.1 Molecular property prediction

The physicochemical properties of organic molecules dictate how they behave in chemical reactions and influence the evolution of organic transformations [50,51]. Therefore, the accurate comprehension and prediction of molecular properties serve as the backbone for the rationality of organic synthesis. As shown in Table 2, molecular properties can be classified into thermodynamic and kinetic parameters, with data sourced from both experiment and computation. Through the discussion of the highlighted ML studies, this section will elaborate on how ML enables quantitative thermodynamic and kinetic property predictions.

(1) Prediction of thermodynamic properties

Thermodynamic properties are inherent characteristics of molecules in an equilibrium state that are used as fundamental parameters to assess the thermodynamics of chemical reactions. Traditional approaches for determining thermodynamic parameters entail both experimental methods and quantum chemical computations, which are accurate but also time- and resource-intensive [110,111]. However, since the thermodynamic property is dictated by the molecular structure, it presents a scenario that is well-suited for ML modeling and prediction. This high-dimensional mapping from molecular structure to thermodynamic property can be learned by data-driven approaches, which could lead to much more accurate and efficient thermodynamic parameter prediction than with conventional techniques. In recent years, significant breakthroughs have been seen in the prediction of various important thermodynamic parameters including pK_a, BDE, and others.

pK_a indicates the degree of proton dissociation from a molecule, which is important to understanding heterolytic X–H bond cleavage energies [112–116] and plays a critical

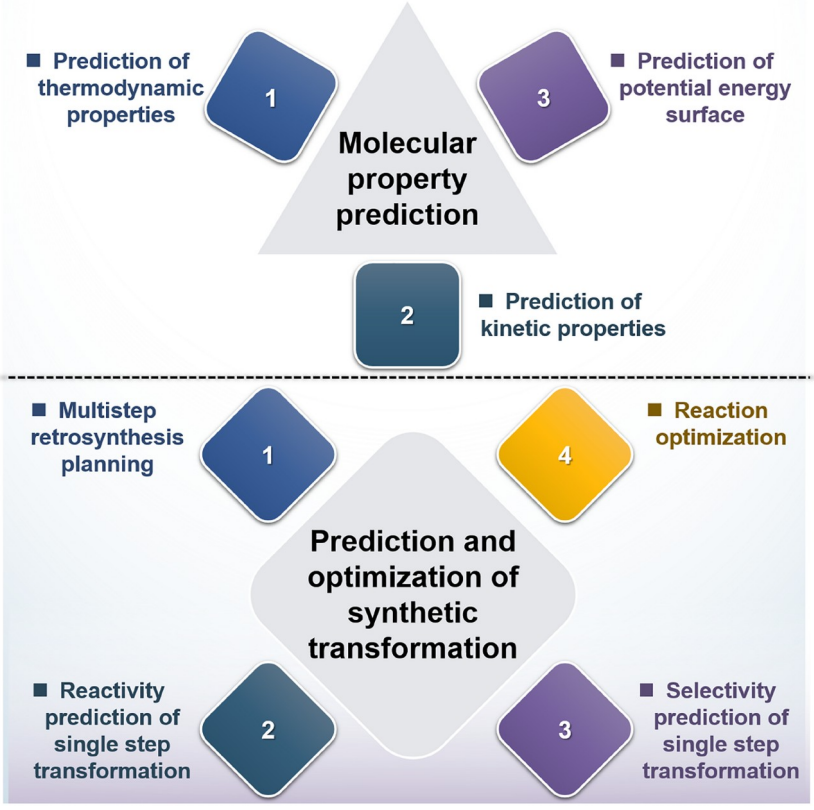

Figure 3 Key directions of AI applications in organic synthesis (color online).

Table 2 Overview of representative molecular properties and models

Category	Property	Data source	Model
Thermodynamics	pK_a	Exp./cal.	Quantum mechanical calculations [52–56], traditional ML methods [21,57–65], GCN [19,66,67], *etc*.
	BDE	Exp./cal.	Theoretical calculations [68–70], quantitative structure-activity relationship (QSAR) [69–75], graph neural network (GNN) [76,77], spectrum-enhanced methods [78], and other ML methods [79–82].
	Quantum chemical properties (HOMO, LUMO, U, H, G, etc.)	Cal.	SchNet [40], PhysNet [83], HMGNN [84], TensorNet [85], ChemRL-GEM [86], Uni-Mol [87], *etc*.
	Physical chemistry-related properties	Exp./cal.	A variety of AI models based on datasets like ESOL [88], FreeSolv [89], Solv@TUM [90], Lipophilicity [91], *etc*.
	Redox potential	Exp./cal.	HOMO/LUMO orbital energy [92], density functional theory (DFT)-calculated descriptors [93], *etc*.
Kinetics	Activation energies	Exp./cal.	MPNN [94] and hybrid reaction models [95]
	Rate constant	Exp.	Gaussian process regression [96]
	nucleophilicity (N) and electrophilicity (E)	Exp./cal.	Physical, topological, quantum chemical descriptors [97–102], GNN [103], *etc*.
	Potential energy surface (PES)	Cal.	Neural networks [104,105], Deep Potential Net [106], CGnets [107], SchNet [40], Δ-machine learned PES [108], stochastic surface walking method [109], *etc*.

role in both chemical and medical sciences [57,114–116]. Although quantum mechanical computations have been extensively applied for pK_a evaluation with high accuracy [54–58], they also suffer from time- and resource-consuming limitations. To achieve the data-driven pK_a prediction, classic ML methods [21,57–65] and graph convolutional neural networks (GCNs) based approaches [66,67] have made tremendous progress. Benefiting from the massive pK_a values in the *i*BonD database, Luo *et al*. [21] have reported an NN-based ML model that can predict the overall pK_a value of a given molecule (macro-pK_a) with an MAE of 0.87 pK_a units in various solvents [20]. For micro-pK_a of a specific X–H bond, Grzybowski *et al*. [66] achieved a mean absolute error (MAE) of 2.1 pK_a units using a GCN model with a DFT-calculated database, which enabled accurate pK_a prediction of a wide range of C–H acids. These strategies have also been extended to the pK_a prediction in protein residues [117,118], which play a crucial role in regulating protein structures and their functions in biological processes.

BDE, which involves the homolysis of chemical bonds, reflects the intrinsic bond strength and is critical in a series of chemical transformations. One representative example is the metal-oxo complex-mediated C–H activation [114], in which the C–H BDEs are closely related to reaction rates. Typically, BDEs could be determined using experimental methods [119] or theoretical calculations with an MAE of around 2 kcal/mol [68–70]. However, these methods, while precise, are costly and inefficient for large-scale analysis. Over the last two decades, early QSAR studies laid the groundwork for efficient and accurate approaches to evaluating BDEs [69–75]. More recent advancements leveraged high-throughput DFT calculations to generate larger BDE datasets with diverse bond types and advanced ML strategies like GNNs [76,77] and spectrum-enhanced strategies [78], offering high-accuracy BDE predictions [79–82]. It is noted that Paton *et al*. [81] have reported an appealing GNN model based on approximately 300k DFT-calculated BDEs, achieving accuracy with an MAE of 0.58 kcal/mol when compared with DFT calculations.

In addition to pK_a and BDE, the ML modeling of a collection of quantum chemical properties (HOMO/LUMO energies, thermal values like *U*, *H*, *G*, *etc*.) of molecules has made significant progress thanks to the creation of the quantum machine (QM)-series database [36,37,120–122]. This large-scale database serves as a powerful data engine that stimulated the development of a series of novel AI frameworks for molecular prediction, including SchNet [40], PhysNet [83], HMGNN [84], TensorNet [85], ChemRL-GEM [86], and Uni-Mol [87]. These models continued to push the state-of-the-art (SOTA) record of molecular property prediction in the QM-series database, achieving an accuracy comparable to DFT calculations. Aside from the synthetic interest, there has been a long-standing interest in predicting molecular properties that are important for drug and material design. Standard databases such as ESOL [88], FreeSolv [89], Solv@TUM [90], and Lipophilicity [91] have been widely used in developing accurate predictive models to aid in the design and screening of drug-like molecules. Redox potential, on the other hand, is an important parameter in electrochemical behavior and has been modeled using a variety of molecular representations, including HOMO/LUMO orbital energy [92] and DFT-calculated descriptors

[93].

(2) Prediction of kinetic properties

In synthetic chemistry, kinetic properties are just as important as thermodynamic properties in determining the practical feasibility of reactions that are theoretically favorable. Their importance extends to critical aspects such as reaction yield, regioselectivity, and stereoselectivity, and they are essential in a variety of processes, including pharmacokinetics [123,124], dynamic kinetic resolution [125], and petroleum cracking [126]. However, the development of kinetic property prediction lags behind that of thermodynamic properties due to the intrinsic complexity and limited data availability [70,127,128].

Reaction rate constants and activation energies are key kinetic properties worthy of ML modeling [129]. In 2020, Green *et al.* [94] utilized a message-passing neural network model with reaction fingerprints to predict activation energies, achieving an MAE of 1.7 kcal/mol. Using transition state modeling, Buttar *et al.* [95] predicted the barrier of nucleophilic aromatic substitution processes with an MAE of 0.77 kcal/mol. For rate constant prediction, Bowman and colleagues [96] employed Gaussian process regression for rate constant prediction for bimolecular chemical reactions. Greaves *et al.* [130] reported a multiple linear regression method to predict the rate constant of the reaction between benzyl bromide and pyridine with an R^2 of 0.92. However, the lack of a substantial rate constant database limits the scope and application of these ML models, highlighting an important subject for future research and data collection efforts.

As fundamental concepts in polar chemistry, nucleophilicity (N) and electrophilicity (E) are quantified with rate constants in specific reactions. Particularly, Mayr *et al.* established the well-known Mayr equation to describe the nucleophilicity and electrophilicity of molecules [131] and then built a database for N/E evaluation in chemical reactions [132]. Several ML modeling attempts have been made based on the empirically known N/E values to forecast the N/E value of novel reagents using physical, topological, or quantum chemical descriptors [97–102] or directly *via* the GNN model [103]. Recently, Luo *et al.* [97] developed a holistic model for predicting both N (R^2 = 0.92, MAE = 1.45) and (R^2 = 0.93, MAE = 1.45) by integrating reactivity structural and physicochemical (rSPOC) descriptors, which was then used to predict the nucleophilicity of a variety of enamine intermediates and NAD(P)H.

(3) Prediction of potential energy surface

Unlike the thermodynamic and kinetic properties, which are characterized by specific numerical values, ML modeling of potential energy surfaces (PESs) requires capturing the continuous relationship between nuclear coordinates and their corresponding energies. The precise prediction of PES is not only theoretically important but also has great practical significance. The AI potential model has received extensive attention in recent years [104,127,133–135], and it can help us understand and simulate molecular systems [136] and synthetic processes [137,138]. Behler and Parrinello made a seminal contribution to this field in 2007 when they used neural networks to create PES at remarkable speeds and efficiency [105]. This was followed by the introduction of Deep Potential Net in 2017 [106], which achieved quantum chemical precision in generating PESs. Further advancements include Schütt *et al.*'s use of SchNet for PES prediction and its application in molecular dynamics simulations of small molecules [40] as well as Clementi *et al.*'s development of CGnets for coarse-grained (CG) molecular modeling [107], extending it to encompass all-atom free energy surfaces in explicit solvation. In recent years, there has been a surge in exciting innovations, such as neural network-based full-dimensional PES constructions for chemical systems [104], Δ-machine learned PES enhancing DFT-based PESs to near CCSD(T) accuracy [108], and the enrichment of the PES library with more comprehensive datasets for chemical systems [139]. These developments have paved the way for an increasing number of modern methods exploring PESs of chemical reactions [140]. One representative example of using PES to elaborate synthetic mechanisms is Liu's study of the mechanism and selectivity of glucose pyrolysis [109]. Using the AI potential model developed by their stochastic surface walking method [141–145], this study detailed an amount of 6,407 elementary reactions and elucidated the mechanistic details and origins of site-selectivity for 5-hydroxymethylfurfural formation.

3.2 Prediction and optimization of synthetic transformation

Because of the massive possibilities of chemical bond cleavage and formation, synthetic transformation inherently poses a multiple-choice question with numerous possible products. Moreover, the issue of synthetic transformation prediction also involves predicting the quantitative outcomes of reactions (yield, selectivity, *etc*.) and strategizing the synthetic pathways for multistep transformations. These problems are highly amenable to data-driven solutions. In fact, even before the advent of modern artificial intelligence technology, the birth and growth of chemoinformatics encompassed the exploration of using data and programming to address these challenges. In recent years, with the accumulation of large-scale synthetic data and the development of advanced reaction modeling frameworks, this field has seen significant progress. On a range of reaction prediction scenarios, AI has demonstrated promising prospects, even offering judgments that surpass those of human chemists.

(1) Multistep retrosynthesis planning

Computer-assisted synthetic planning (CASP), particu-

larly the strategic planning of multi-step retrosynthesis, represents one of the oldest yet most vibrant challenges in AI synthetic chemistry [146]. The crux of the challenge in retrosynthesis planning lies in the construction of a coherent and reasonable multi-step synthetic network, followed by the execution of an efficient and rational search and scoring process within this network. Finally, the synthesis pathways must be regressed to the available building blocks, ensuring a practical and feasible approach to the synthesis design. To realize the multistep retrosynthesis, a series of innovative algorithm designs have been proposed in recent years. In this regard, Segler *et al.* [147,148] utilized the Monte Carlo Tree Search (MCTS) algorithm to devise synthetic routes for small organic molecules. Kishimoto *et al*. [149] introduced the DFPN-E method, integrating depth-first proof-number search (DFPN) with heuristic edge initialization, showcasing a time advantage over the MCTS algorithm with comparable success rates. Chen *et al.* [150] presented Retro*, a neural-based A*-like algorithm, utilizing an AND-OR search tree and an optimal priority search strategy, offering a more efficient approach to searching reaction pathways. Xie *et al.* [151] presented a graph-based search algorithm called RetroGraph, further enhancing the performance of A*-like search algorithms to reduce molecular redundancy in tree-based search methods. Kim *et al.* [152] introduced Retro*+, a self-improving framework training a single-step model to emulate successful trajectories, maximizing success rates and leveraging simulated experiences for model enhancement. Yu *et al.* [153] proposed GRASP, a goal-driven actor-critic method, utilized for seeking routes with specific predefined objectives, such as building block materials. Recently, Liu *et al.* [154] presented PDVN, a dual-value network planning, constructing two distinct value networks to predict synthesizability and cost, enhancing search success rates, optimizing model invocations, and aiding in identifying shorter synthetic routes.

With the above algorithm advancements, reports on computer-aided multi-step route design in chemical synthesis are emerging. The Jensen group [155] employed ASKCOS for multi-step retrosynthetic route design of 15 drug molecules, including (*S*)-warfarin and safinamide, and validated the synthetic feasibility through a robotic flow chemistry platform. The Grzybowski team [12] demonstrated SYNTHIA's powerful retrosynthetic design capabilities, passing the Turing test in which chemists cannot differentiate the AI-designed and the human-designed synthetic routes for the studied compounds. A series of SYNTHIA-predicted routes for natural products are experimentally executed, including challenging targets of (−)-Dauricine, Tacamonidine, and Lamellodysidine A (Figure 4a). It is worth noticing that Tacamonidine and Lamellodysidine A were synthesized for the first time. The Cernak research group [156] utilized SYNTHIA to study 12 potential anti-COVID-19 drugs, experimentally validating four predicted routes for umifenovir and one predicted route for bromhexine, highlighting that automated retrosynthetic predictions can rapidly identify alternative starting material supply chains for pharmaceuticals. Cernak and colleagues [45] further demonstrated that human chemists can harness the heuristic value of AI predictions to achieve out-of-box synthetic innovations. They group-employed SYNTHIA for the retrosynthetic route prediction of (−)-stemoamide. Based on the myriad of SYNTHIA-predicted pathways, they proposed a concept of graph edit distance to quantify the synthetic impact of AI-suggested single-step transformations (Figure 4b). Through this, they were able to realize a remarkable 3-step synthesis of (−)-stemoamide (Figure 4c). Interestingly, the AI tool for retrosynthesis analysis can also be applied in a reversed fashion to guide forward synthetic possibilities. The Grzybowski team [157] utilized the forward-synthesis Allchemy platform to generate a plethora of synthetic networks from approximately 200 commercially recovered waste chemicals. They selected numerous viable synthetic routes and experimentally validated several of them. The continuous reporting of computer-aided multi-step reaction planning, encompassing both algorithm development and experimental applications, underscores the growing significance of this field in the work of synthetic chemists.

(2) Reactivity prediction of single step transformation

For molecular synthesis, although many reactions appear theoretically feasible, the reactions that can actually achieve uniformly high reactivity and selectivity are exceptionally rare [158–160]. More commonly, most reactions only achieve the desired efficiency and selectivity under a delicate combination of substrate, catalyst, and conditions. Therefore, accurate evaluation of the reactivity and selectivity of single-step transformation is equally crucial for the successful design of molecular synthesis [161,162]. However, due to the vast molecular structural space and the multitude of controlling factors, there is no simple formulaic equation capable of quantitatively describing the universal laws of molecular synthesis. The QSAR of single-step transformation still remains one of the core challenges in AI synthesis [163–165]. Facing this challenge, traditional research has typically relied on experience-driven strategies: by summarizing the available data, synthetic chemists are able to derive a local structure-activity relationship for the specific target, which is then used for the rational design and improvement of synthetic transformation. However, this empirical approach lacks precision and predictive power, and conflicting rules can exist. This situation makes random selection and trial-and-error inevitable in designing and screening actual synthetic explorations.

Recently, data-driven approaches have brought a new perspective to solve the problem of single-step QSAR prediction [127,166–168]. Benefiting from advanced AI algo-

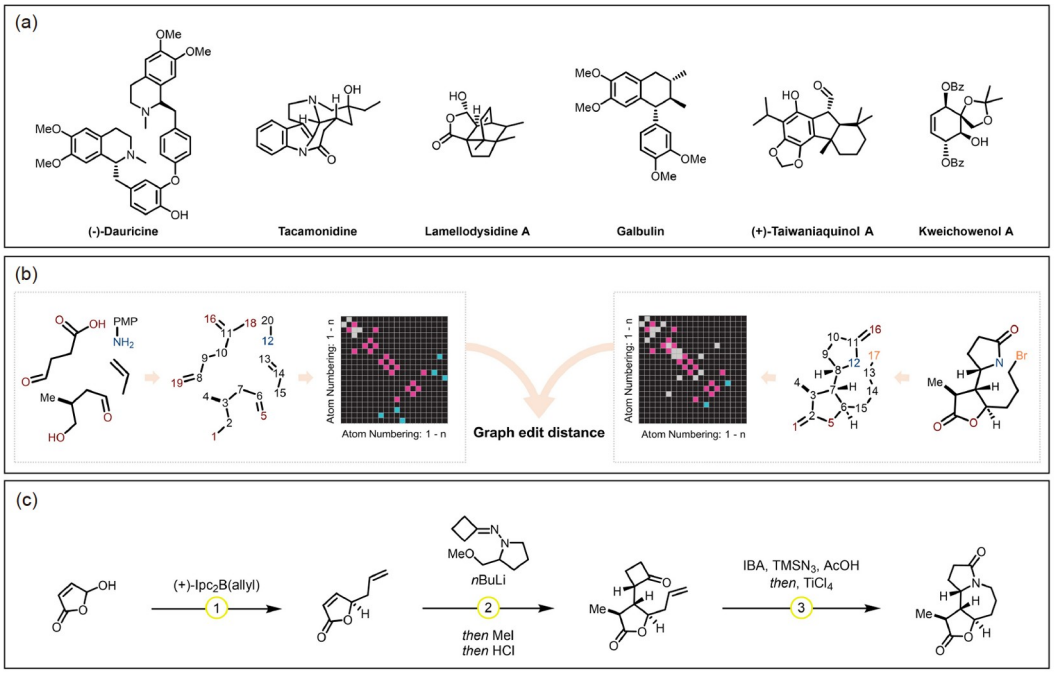

Figure 4 Representative applications of SYNTHIA. (a) Highlighted complex organic molecules whose SYNTHIA's predicted synthetic routes have been experimentally verified. (b) Calculation of the graph edit distance between two synthetic intermediates based on bond connections. Reproduced with permission from Ref. [45]. Copyright 2023, American Association for the Advancement of Science. (c) 3-step synthesis of (−)-stemoamide inspired by SYNTHIA (color online).

rithms and rich chemical data, a series of studies have shown that ML models are able to accurately predict reaction yields and selectivities [169–175], even surpassing the judgment of experienced chemists in some cases [12,176]. More importantly, these models can assist chemists in efficiently screening new catalysts for target reactions [177–179], providing powerful AI tools for molecular synthesis. These studies revealed the remarkable potential of ML technology in synthetic chemistry, promising to accelerate the process from the development of synthetic methods to the discovery of functional molecules.

For the palladium-catalyzed Buchwald-Hartwig cross-coupling reactions, Doyle *et al*. [180] demonstrated the potential of ML in predicting reaction yields. Utilizing a high-throughput synthetic platform, they reliably evaluated the yields of 4,140 reactions comprising a diverse range of substrates, catalysts, additives, and bases (Figure 5a). Through quantum chemistry computations and customized scripts, a series of atomic, molecular, and vibrational descriptors were automatically generated. Employing a random forest regression algorithm, they achieved an R^2 of 0.92 and a root mean square error (RMSE) of 7.8% across a 70% (training)/30% (validation) data split. Furthermore, the trained ML model is able to predict the outcomes for unseen additives, showcasing the extrapolative predictive power of the established yield model. Interestingly, the model interpretation revealed that the descriptors of isoxazoles were crucial for yield predictions. Following this mechanistic hint, the authors subsequently discovered that the active isoxazoles were able to inhibit palladium's catalytic activity through oxidative addition. This study, combining high-throughput experimentation with ML modeling, unveiled the attractive potential of data-driven research paradigm for reaction design and screening. It provides a powerful AI tool for evaluating the productivity of Buchwald-Hartwig cross-coupling reactions and enriches the understanding of the reaction mechanism.

By merging automation and ML modeling, Liao and colleagues [181] achieved selective Pd-catalyzed functionalization of sterically hindered aromatic *meta*-C–H bonds. They employed a synergistic protocol combining photoinduced C–H carboxylation, carboxy-directed Pd-catalyzed C–H functionalization, and microwave-assisted decarboxylation, using CO_2 as a traceless director for targeted *meta* C–H functionalization (Figure 5b). Through high-throughput experiments, they efficiently executed 1,032 reactions to explore a remarkable substrate scope, thereby providing comprehensive insights into the reaction's synthetic potential. With this dataset in hand, they developed a yield prediction model using a message-passing neural network with pre-training from the USPTO dataset, which achieved an R^2 of 0.750 and an MAE of 7.2% in 5-fold cross-validation. In addition, this model is able to accurately predict the reaction outcome for unseen substrates, demonstrating significant

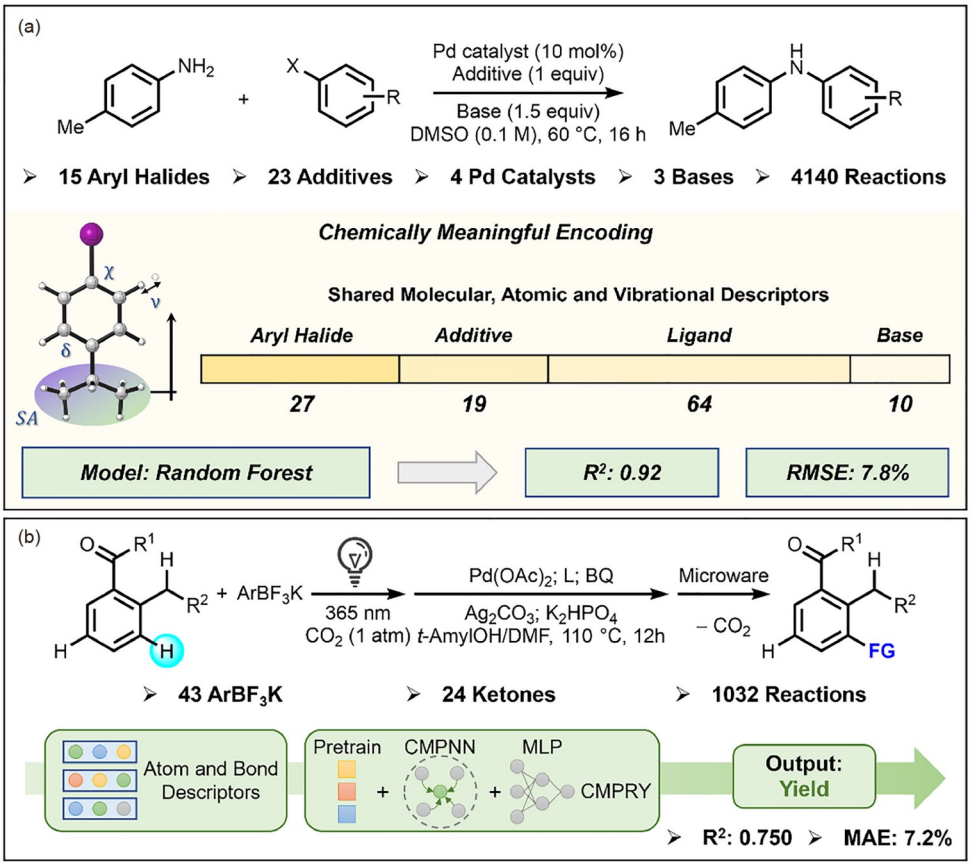

Figure 5 Selected ML yield prediction studies of organic transformation. (a) ML approach and performance of yield prediction of Pd-catalyzed Buchwald-Hartwig cross-coupling reactions. (b) ML approach and performance of yield prediction of Pd-catalyzed functionalization of sterically hindered aromatic *meta*-C–H bonds (color online).

advantages of high-throughput experimentation and ML-assisted yield prediction in exploring novel synthetic reactions.

Differing from the complete HTE dataset of chemical spaces, the synthetic exploration in real-world applications tends to be sparse and significantly more diverse in molecular selections. To investigate whether ML models could meet the challenge of predicting such scenarios, Wiest and colleagues [182] extracted and processed the data from AstraZeneca's ELNs, creating a dataset for Buchwald-Hartwig reactions. This dataset included 781 reactions involving 340 aromatic halides, 260 amines, 24 ligands, 15 bases, and 15 solvents. Moreover, it contained a substantial number of low-yield or nonproductive reactions, with 39.9% yielding no product. This ELN dataset reflected the reality of synthetic transformation in pharmaceutical applications, presenting a significant challenge for ML modeling. The authors found that all attempted models, including the classic regression algorithms using RDKit features and more advanced Yield-BERT [183] model, failed to provide meaningful predictions, with the best model achieving an R^2 of only 0.266. This finding, contrasting with the success of similar models on HTE datasets, indicated that ML models still need significant improvement in handling real-world synthetic scenarios and highlighted the need for caution in modeling with legacy yield data.

Also targeting the challenge of biased distributions in literature data, Glorius *et al*. [184] noticed the importance of negative data in structure-activity relationship modeling. They found that, despite having up to 190,000 yield data from literature, models still struggled to achieve reliable predictions. Whether using traditional modeling methods or Yield-BERT [183] models, none could provide meaningful regression results, which is consistent with the above study from Wiest. By manipulating the data extraction and including additional random noise, the Glorius group attributed these poor modeling results primarily to the bias in literature data rather than the noise in experimental data. This bias stemmed from both selective sampling of reaction spaces in literature reports and a tendency to publish positive reaction outcomes. To address this issue, the authors proposed two strategies: purposefully conducting additional experiments to gather data on low-performing synthetic space, and data augmentation to help mitigate the issues caused by overly biased sampling. This work further revealed the importance of data distribution in synthetic chemistry modeling, high-

lighting the critical need for comprehensive, diversified, and fair evaluations and reporting in synthetic investigations.

(3) Selectivity prediction of single-step transformation

As a key component of structure-performance relationships, selectivity is also a crucial target for ML predictions in synthetic chemistry [161,164,166,185,186]. To realize the data-driven prediction of asymmetric catalysis, the Denmark group [187] reported the successful ML application in BINOL phosphoric acid (BPA)-catalyzed asymmetric addition of imines (Figure 6a), demonstrating the advantages of AI in solving stereoselectivity problems. The authors introduced innovative designs in both data selection and ML modeling. For data selection, they proposed a concept called universal training set (UTS), employing the Kennard-Stone algorithm to select chemically representative substances within a space composed of steric and electronic descriptors. This selection method, independent of reaction and mechanistic understanding, is solely based on the physicochemical properties of studied molecules, thereby rendering the chosen BPA set broadly applicable for modeling of BPA-involved transformations. In addition, to accurately characterize the complex steric environment of BPA molecules, a novel stereochemical descriptor called "average steric occupancy" (ASO) was developed. This descriptor is based on molecular occupancy at grid points within a cubic lattice: for each grid point, a value of 0 or 1 is assigned based on whether molecules occupy this position. The grid values were subsequently averaged across all conformers to generate the uniformed, high-dimensional ASO descriptor representing the steric environment of the molecule. Integrating these innovative designs, the authors attempted ML predictions of enantioselectivity on a dataset comprising 43 BPAs, 5 imines, and 5 thiols, totaling 1,075 reactions. The constructed neural network model precisely predicted the target enantioselectivity, with a mean absolute deviation (MAD) of about 0.15 kcal/mol. Moreover, the model's reliability was validated in several out-of-sample and out-of-range tasks, accurately predicting unseen catalysts and successfully differentiating superior ones. This work, with the innovative workflow and descriptor designs, provides a key reference for data-driven modeling of stereoselectivity, highlighting the potential of AI technology in addressing asymmetric challenges.

Interestingly, Sigman and colleagues [188] also explored the stereoselectivity prediction of BPA-catalyzed imine addition reactions from the perspective of multivariate linear

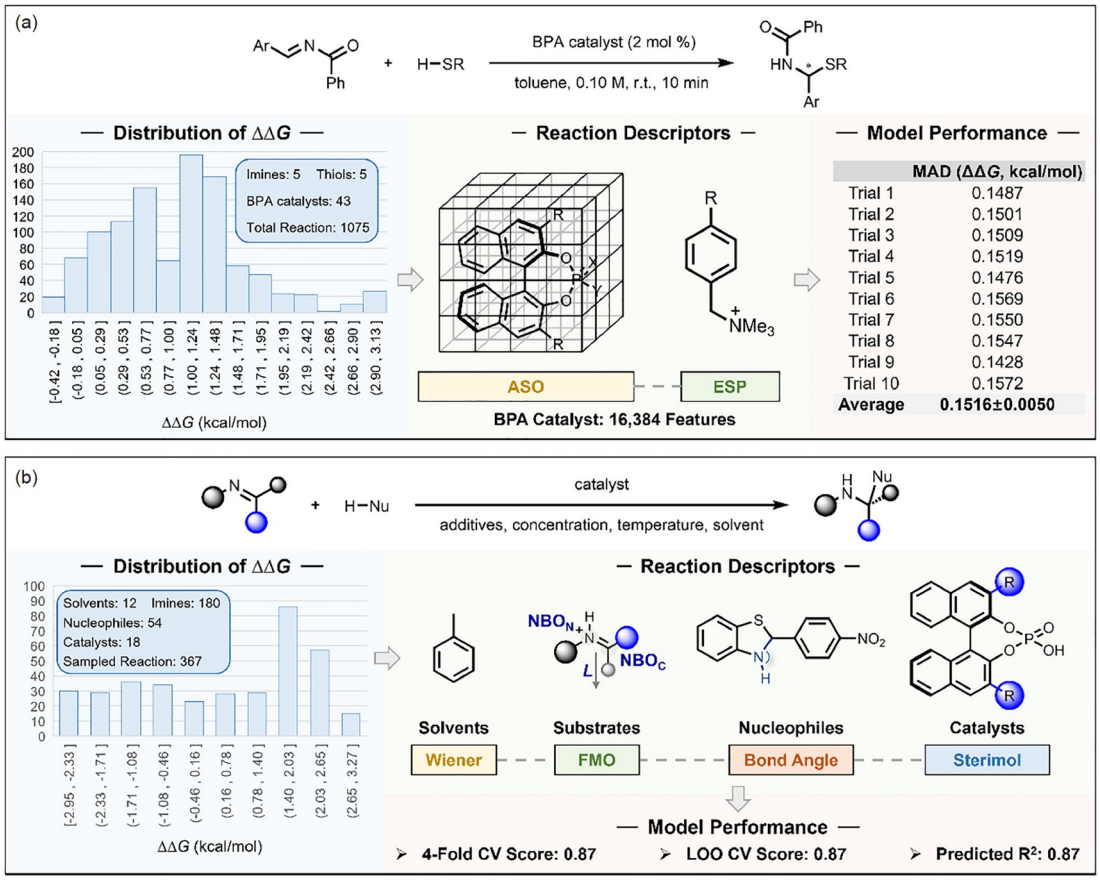

Figure 6 Selected ML enantioselectivity prediction studies of chiral phosphoric acid-catalyzed imine addition. (a) Data distribution, molecular descriptors with the design of ASO, and model performances. (b) Data distribution, physical organic descriptors, and the model performances using the multivariate linear regression approach (color online).

regression using physical organic parameters (Figure 6b). They posited that by uncovering the common mechanistic features of all reaction components, a comprehensive understanding of the factors controlling the reactivity and selectivity could be achieved. With this comprehensive set of physical organic controlling factors and leveraging statistical modeling, predictions for different structural motifs within a single model became feasible. For the studied BPA catalysis, the authors systemically parameterized the involved reaction components and catalysts from a physical organic chemistry standpoint, leading to the compilation of 313 parameters expressing steric and electronic effects. Based on this, multivariate linear regressions were made on an enantioselectivity dataset of 367 reactions compiled from relevant literature. The linear model achieved the remarkable performance of R^2 close to 0.9. Notably, when applied to the exact same dataset published by Denmark [187], Sigman's model also achieved excellent predictive performance, accurately identifying the selective catalysts. This highlights that the mechanism-based physical organic chemistry parameters can enable powerful QSAR prediction of focal datasets without the usage of sophisticated regression algorithms, which provides an alternative approach for quantitative predictions of stereoselectivity.

Exciting advances in stereoselectivity prediction were also made for transition metal catalysis [165,189,190]. Focusing on the asymmetric hydrogenation of olefins [189], Hong and colleagues [191] reported a productive ML modeling study. Due to the highly sparse and biased nature of the selected datasets from literature, a novel modeling strategy called "hierarchical learning" was proposed to overcome the bias and achieve extrapolative prediction. The core idea is to view the structure-selectivity relationship as a superposition of a universal relationship and local perturbations. A base model representing the universal structure-activity relationship is learned through the representative data samplings, followed by training a delta model with the neighboring data close to the target reaction, thus learning the perturbations of the relationship. The superposition of layered models yields the final prediction, which can be considered as an approach to transfer learning. This transfer learning strategy achieved excellent prediction in the asymmetric hydrogenation of olefins, requiring only limited reaction data of the target alkene substrate for satisfying modeling. The collaboration between Hong and Ackermann further applied the hierarchical learning strategy to explore the holistic synthetic space of electrochemical Pd-catalyzed C–H alkenylation (Figure 7a) [192], systematically studying the enantioselectivities of 846,720 reaction combinations, which demonstrated the appealing advantages of data-driven method in achieving the comprehensive knowledge of synthetic space. They also utilized this transfer learning protocol in virtual catalyst screening for asymmetric Co-catalyzed C–H alkenylation, which successfully predicted and verified an intriguing chiral carboxylic acid with excellent enantioselectivity [193].

In addition to enantioselectivity, ML modeling has also been applied to other categories of stereoselectivity [172,194,195]. One representative study is Grzybowski's work [194] in Diels-Alder reaction (Figure 7b). Using physical organic descriptors, the authors successfully predicted the major regio-, site-, and diastereoisomers using ML modeling. They found that using physical organic descriptors, as opposed to naive molecular fingerprints, significantly improved the model performance. By capturing electronic effects with Hammett constants and steric properties with TSEI indices, the chemical descriptors combined with a random forest classifier achieved an excellent prediction accuracy for regio- (93.6%), site- (91.3%), and diastereoselectivities (89.2%). The authors further demonstrated that the prediction performance of the Hammett-TSEI-based random forest classifier received less effect by the dataset partitioning compared to other encodings, indicating that the physical organic descriptors can enable the model to learn the organic structure-performance relationship and accurately predict outcomes for compounds unseen during model training.

Regioselectivity prediction has also been realized using ML methods, as evidenced by a series of exciting advances in recent years [194,196–199]. The Hong group [197] has conducted ML studies on the regioselectivity of radical C–H functionalization of arenes (Figure 8a). Based on previous mechanistic understandings [200], they systematically computed the DFT barriers of the rate-determining step for a myriad of substrates, obtaining regioselectivity data for 9,438 reactions. They found that the physical organic descriptor, with only a few dozen dimensions, achieved satisfying regression results comparable to other typical higher-dimensional descriptors (smooth overlap of atomic positions (SOAP), atom-centered symmetry functions (ACSF), *etc.*). The trained random forest model accurately predicted the reaction sites with 94.2% accuracy and determined the degree of selectivity with 89.9% accuracy. Subsequently, the model's predictions were compared with reported experimental results on complex polysubstituted aromatics. Despite being trained only on DFT data and not having been exposed to the experimental complex compounds, the model still performed with convincing accuracy. This work not only demonstrates that the DFT computation can provide a reliable data source for ML modeling of selectivity problems and can be directly applied to experimental prediction and verification, but emphasizes the importance of local physical organic descriptors of reaction sites in regioselectivity modeling.

Jensen and colleagues [198] advanced the prediction modeling of regioselectivity for synthetic transformations,

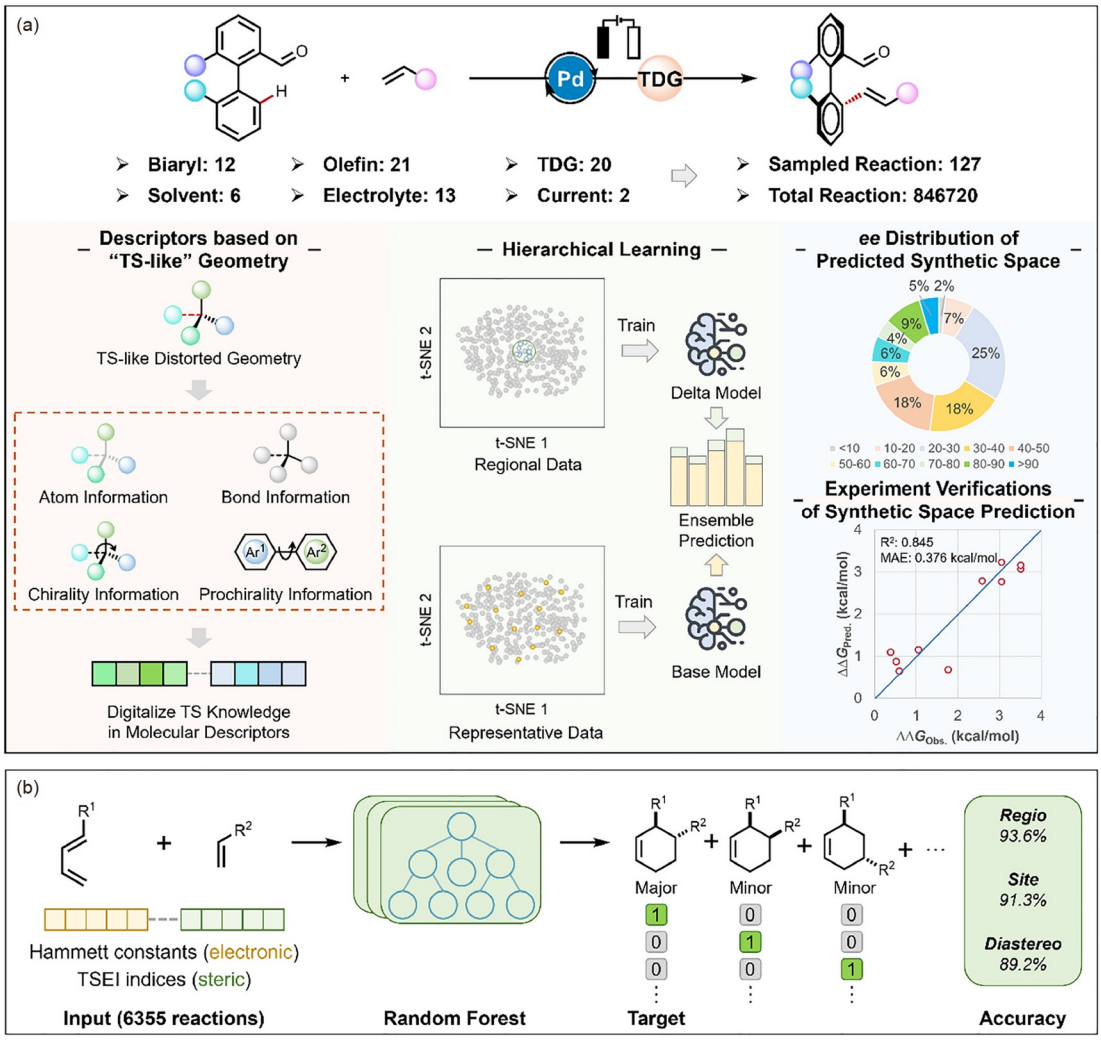

Figure 7 Selected ML stereoselectivity prediction studies of Pd-catalyzed C–H alkenylation and Diels-Alder reaction. (a) Descriptor design, ML approach, and the model performances and predictions of the enantioselectivities of pallada-electrocatalyzed C–H activation. (b) Descriptor design, ML approach, and the model performances of the regio-, site-, and diastereoselectivities of Diels-Alder reaction (color online).

including aromatic C–H functionalization and C–X substitution. In their work, the implicit molecular representations derived from ML were combined with explicit quantum chemical properties (Figure 8b). The machine-learned molecular representation was realized by a GNN based on the Weisfeiler-Lehman network framework. Quantum chemical descriptors of organic molecules (atomic charges, Fukui indices, nuclear magnetic resonance (NMR) shielding constants, *etc*.) were calculated at the B3LYP/def2-SVP level using GFN2-xTB optimized structures. By combining these two types of molecular representations, they developed an ML model that accurately predicts the regioselectivity of the target reactions. To circumvent the time- and resource-consuming quantum chemical calculations, they demonstrated the appealing potential of combining multiple ML models. For this, they trained a directed message-passing neural network to predict QM properties of molecules, using these predictions instead of DFT-computed parameters as input for the selectivity model. This strategy provided an end-to-end model that can predict selectivity from SMILES within milliseconds. The evaluation showed that this fusion model achieved an accuracy of 89.7% for aromatic C–H functionalization reactions, 96.7% for aromatic C–X substitution reactions, and 97.2% for other substitution reactions. This work illustrates the complementary nature of data-driven representation and quantum chemical parameters for molecular encoding, while also highlighting the potential of on-the-fly quantum chemical property prediction and derived reactivity/selectivity modeling in synthetic chemistry.

Hartwig and colleagues [201] successfully merged expert rules with data modeling to predict the regioselectivity of Ir-catalyzed C–H borylation reactions (Figure 8c). They combined literature results with specifically sampled low-selectivity data and representative intermolecular competition

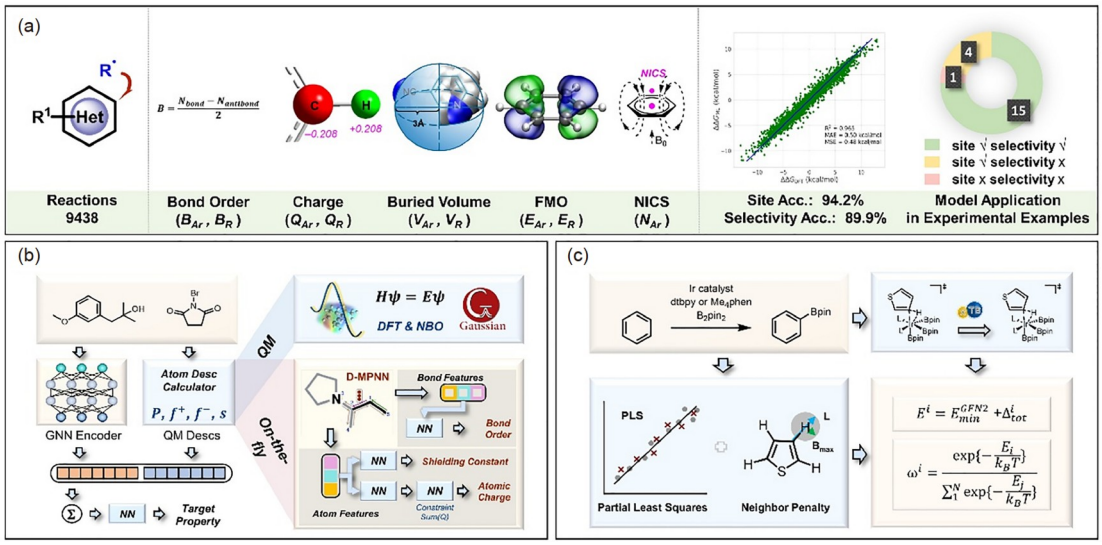

Figure 8 Selected ML regioselectivity prediction studies of organic transformation. (a) Descriptor design and ML performances of radical C–H functionalization of arenes. Reproduced with permission from Ref. [197]. Copyright 2020, John Wiley & Sons. (b) Regioselectivity prediction model of aromatic C–H functionalization and C–X substitution that combines the machine-learned representation by GNN and the calculated atomic descriptors. Reproduced with permission from Ref. [198]. Copyright 2021, Royal Society of Chemistry. (c) Regioselectivity prediction model of Ir-catalyzed C–H borylation that combines the xTB calculation, the ML regression, and the expert rule for neighboring substituent influence (color online).

experiments, forming a data set for regioselectivity training. In selectivity modeling, they combined computational chemistry, data-driven approaches, and expert experience. Starting with a rough estimation of reaction activation barriers using the xTB method, they combined it with a partial least squares (PLS) model to predict the regioselectivity. Considering the limited scope of modeling data, they further implemented an expert rule to express the influence of neighboring substituents. Combining model predictions with rule-based corrections, they provided a predictive model for evaluating borylation sites. This approach maximized the benefits of computation, modeling, and expert experience, resulting in a highly effective selectivity prediction model. The model's predictions aligned well with experimental verification across various compounds, which was also compared with predictions from experienced human chemists, outperforming them in tests on a few showcase complex compounds. This study demonstrates the complementary nature of human expertise and ML modeling, showing how expert rules can enhance and improve the predictive capabilities of ML models.

(4) Reaction optimization

Data-driven reaction optimization allows an effective strategy to accelerate the discovery and improvement of synthetic processes by providing actionable recommendations for reaction conditions. This is an optimization problem in a defined chemical space, which does not fall into the category of regression or classification, necessitating an iterative workflow with experimental testing for correct and complete predictions in the face of chemical nuances. Traditionally, reaction optimization relies on chemists' knowledge and design of experiment (DoE) methods, which, though automatable, do not harness the statistical value of the accumulated data and demand significant experiment efforts. In addition, the efficacy of these models relies heavily on the quality of the training data. Beker *et al.* [202] argued that noise and bias in literature data could hinder the creation of models that surpass literature popularity trends, emphasizing the importance of high-quality training data.

In 2018, the Aspuru-Guzik group [203] introduced Phoenics, an algorithm using Bayesian optimization for global optimization in chemical experimentation. Phoenics proposes new conditions based on previous observations, efficiently identifying optimal conditions and demonstrating applicability in complex case studies like the Oregonator, a nonlinear chemical reaction network. Sunoj and colleagues [204] developed an ML model for discovering catalysts in asymmetric hydrogenation, accurately predicting enantiomeric excess with an RMSE of 8.4 ± 1.8 using molecular parameters from 368 substrate-catalyst combinations. The model successfully predicted out-of-sample data, indicating potential for catalyst discovery and substrate selection. In 2021, Doyle, Adams, and colleagues [176] reported a Bayesian reaction optimization framework integrating algorithms into daily lab practices. They applied Bayesian optimization to Mitsunobu and deoxyfluorination reactions, enabling more efficient synthesis of value-added compounds through data-driven experimental decisions. Wang and colleagues [205] used ML to accelerate Cu catalyst discovery and optimization for CO_2 reduction, identifying critical features and facilitating catalyst design and validation. In addition, the ML-assisted reaction optimiza-

tion extends to materials. Norquist *et al*. [206] used support vector machine (SVM) for an ML-assisted materials discovery from failed experiments. Cooper's group [207] integrated robotic experimentation and high-throughput computation to explore high-activity linear polymers for hydrogen evolution photocatalysts. It is also demonstrated that the ML model can recommend new reactions and generate hypotheses about crystal formation, emphasizing its versatility beyond traditional organic synthesis.

4 AI applications in polymer synthesis

As key substances in the fields of molecular science and functional materials, the AI application in polymer synthesis has received wide interest and significant progress in recent years. Unlike small molecules with well-defined structures, polymers' continuous molecular structure adds more possibilities in terms of structure and functionality. However, this also brings new challenges for data modeling. These challenges manifest both in how to represent the molecular structure of polymers in a data-driven format for ML models and in how to rationally establish quantitative relationships between the polymer synthesis process and the structure/properties of the generated polymers. Furthermore, biomacromolecules like proteins and DNA are naturally programmable macromolecular machines. Addressing the synthesis of such biomacromolecules through data-driven solutions is not only a focal point of synthetic interest but also forms a key component in the field of bioinformatics. This section delves into the research on AI applications in polymer synthesis, with key directions highlighted in Figure 9. It aims to reveal the new opportunities AI brings to this field, as illustrated through discussions on representative works.

4.1 Structure-property relationship prediction of polymer

Polymers constitute a significant class of materials that are ubiquitous, with applications spanning from daily products, including plastics and rubbers, to cutting-edge high-tech products in electronics, photonics, and biomedicines [208]. The highly tunable functionality of polymers arises from their remarkable diversity at both microscales (*e.g.*, chemical composition, atomic-level connectivity) and macroscales (*e.g.*, crystallinity, phase separation) [209]. Nevertheless, the vast and complex chemical and morphological spaces hinder the discovery of novel polymeric materials for specific purposes.

Materials science has witnessed transformative advancements through the integration of ML into polymer property prediction. ML techniques are able to leverage pre-existing experimental data and computational data from first-principles calculations or molecular dynamic simulations, establishing models for rapid and accurate predictions of the properties of new polymer materials [210]. The initial step in modeling the structure-property prediction of polymers involves defining their representation at atomic/molecular levels [211]. However, traditional text-based representations are labor-intensive, computationally demanding, and lack adaptability to diverse polymer classes. These drawbacks hinder the development of AI/ML pipelines for high-throughput applications. In parallel, it is also non-trivial to efficiently probe existing databases for further discoveries. Advanced ML frameworks provide promising prospects for bridging the gap by exploiting available data.

As illustrated in Figure 10a, graph-based representations, which directly capture topological information of chemical structures, show favorability for property prediction tasks of polymers. Simine *et al*. [212] predicted ultraviolet-visible (UV-vis) spectroscopy of conjugated polymers directly from CG representations *via* a deep-learning model of long-short-term memory recurrent neural network. The approach demonstrated the potential to investigate organic optoelectronics through computational experiments invoking CG representations. Wang and colleagues [213] utilized CG representations to construct a high-dimensional design space. Bayesian optimization process efficiently explored this continuous space, offering comprehensive insights into molecular-level relationships influencing the lithium conductivity of polymer electrolytes. More recently, Aldeghi *et al*. [214] introduced a graph representation of molecular ensembles that captured key features including monomer compositions and chain architectures, using a weighted directed message-passing neural network tailored for polymer property prediction. The platform established a database of over 40,000 possible copolymers *via* calculation of electron affinity and ionization potential and achieved superior accuracy than off-the-shelf material informatics methods.

The Ramprasad group [209,215] constituted a user-friendly structure-property prediction platform named "Polymer Genome". The informatics platform leveraged three hierarchical levels of fingerprints to capture features critical to describe a specific polymer property, which spanned from three-atom fragments, descriptors of the quantitative structure-property relationship type, to morphological descriptors such as the fraction of side-chain atoms (Figure 10b) [208]. Researchers utilized ML algorithms based on Gaussian process regression to generate prediction models that were implemented in the online platform "Polymer Genome".

Fingerprints as inputs for predictive models tend to attain the relatively best performance [216]. However, hierarchical handcrafted fingerprints, necessitating chemical intuition, consume an amount of time due to complicated computa-

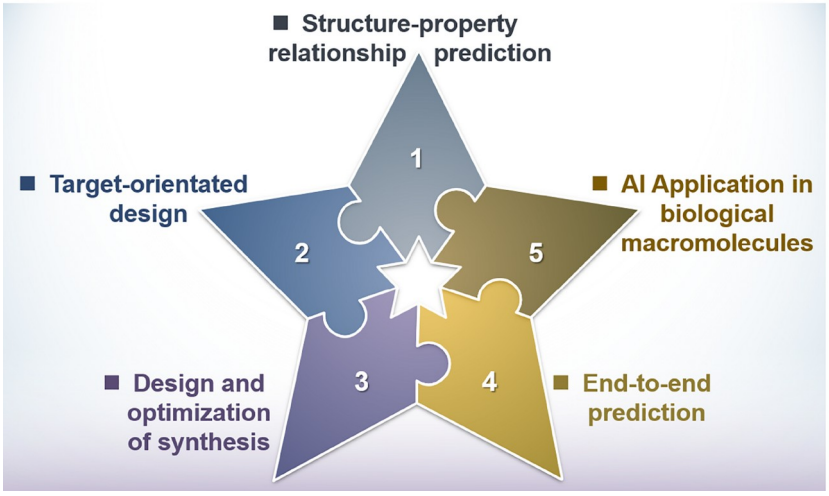

Figure 9 Key directions of AI applications in polymer synthesis (color online).

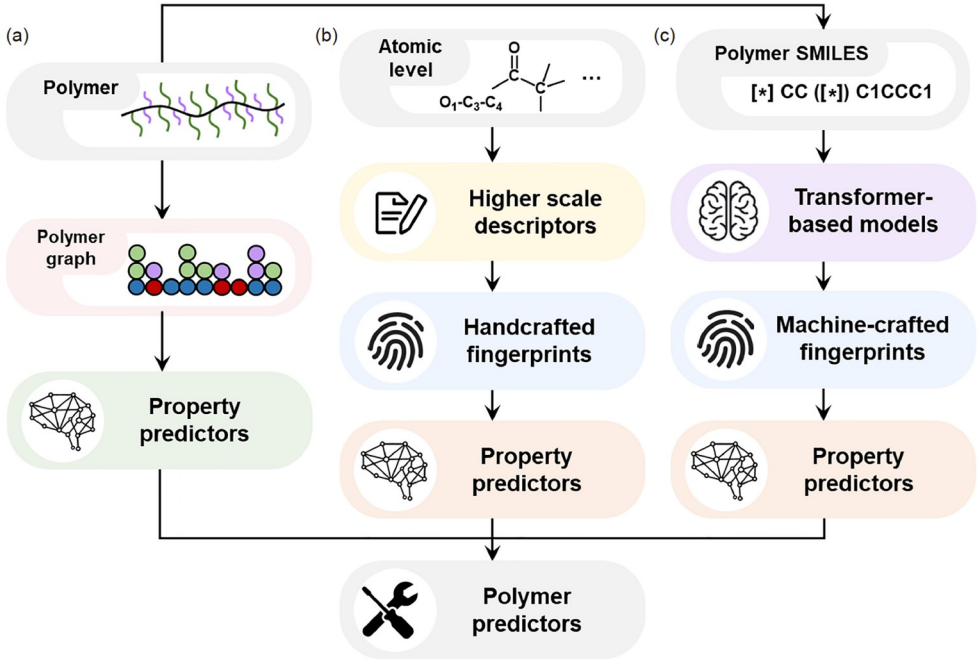

Figure 10 Illustration of (a) graph-based representations and (b) Transformer-based models favorable for structure-property relationship prediction of polymers (color online).

tions for model training and inference. Recent advancements in NLP have established Transformer as a powerful AI framework for language modeling. SMILES strings, considered the "chemical language" of polymers, entitle Transformer-based models to the opportunity for application in polymer science. Xu and colleagues [217] introduced TransPolymer, a Transformer-based model benefitting from pretraining on a large unlabeled dataset. This model showcased the importance of chemical awareness in modeling polymer sequences, affording a robust tool for structural-property relationship exploration. Similarly, Kuenneth *et al.* [218] trained polyBERT on 100 million polymer SMILES strings of hypothetical polymers to function as a chemical linguist. Integrated into multitask deep neural networks, the fully machine-learned polyBERT fingerprints predicted polymer properties at unparalleled speed with unimpaired accuracy, surpassing the SOTA record of handcrafted Polymer Genome fingerprints (Figure 10c).

Deep-learning architectures have revolutionized structure-property prediction of polymers by automatically learning expressive representations (Figure 11a). Rahman *et al.* [219] proposed a CNN-based framework that predicted the critical mechanical property, namely pullout force, of carbon nanotube-polymer interfaces. Park *et al.* [220] utilized GCNs for

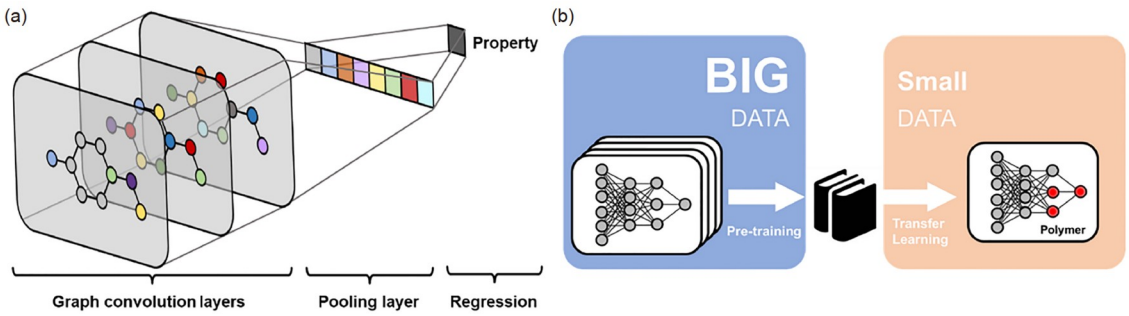

Figure 11 Schematic representation of (a) deep-learning architectures and (b) transfer learning-based frameworks for polymer property prediction (color online).

predicting the thermal and mechanical properties of polymers. They found that GCNs, especially when combined with neural network regression, could slightly outperform the widely used extended-connectivity circular fingerprint (ECFP) representation.

In addition to the structure-property relationship prediction, correlations between chemical, electronic, mechanical, and thermodynamic properties offer alternative avenues for effective property prediction models. Transfer learning leveraging models based on interrelated properties proves promising for predicting target properties even with minimal data (Figure 11b). The Yoshida group [221] developed XenonPy.MDL, a pretrained model library with over 140,000 models for diverse properties of organic small molecules, polymers, and inorganic materials. Their frameworks, exemplified by neural network models, demonstrated efficient property prediction for extremely small datasets. Through a multi-fidelity fusion strategy that addresses the limitation of experimental data in quantity and diversity, the Ramprasad group [222] trained the model on the low-fidelity but abundant data set employing group contribution methods to predict polymer crystallinity under high-fidelity accuracy. Later, the same group [223] advocated multi-task learning by exploiting intercorrelations between various property datasets, yielding efficient, scalable, and interpretable models for polymer property prediction.

4.2 Target-orientated design of polymer

In addition to the forward structure-property relationship prediction, it would be ideal to inverse design the desired polymer with target property. This concept of inverse design presents a novel paradigm, departing from the traditional Edisonian method, which accompanies time- and labor-intensive exploration reliant on human intuition with inherent biases and knowledge limitations. This approach enables generating polymers with superior functionality or properties by navigating the chemical space informed by data-driven strategies. Two avenues, high-throughput screening and advanced ML algorithms, are recognized as pivotal protocols to achieve the target-oriented polymer design (Figure 12).

In the context of high-throughput screening, researchers should narrow the chemical space by defining the inputs of polymer fragments and adjoining rules based on their prior knowledge and chemical intuition, which would simultaneously ensure the validation of combining building blocks. For example, the Ramprasad group leveraged a polymer database derived from first principles, exploring the linear combination of 7 basic building blocks to recommend novel dielectric polymers [224]. In a similar manner, Afzal *et al.* [225] identified polyimides with exceptional refractive index values *via* high-throughput virtual screening, which could access a massive library of polyimide structures composed of 29 building blocks. However, integrating polymer fragments as inputs is prone to neglecting interaction between polymer chains and other influential factors in realistic production.

Active learning and the derived AI-driven space exploration have also been utilized in the search for polymer candidates. Bayesian optimization, a noise-tolerant and global optimization strategy free from assumptions of functional forms, has also been implemented in polymer design. The workflow utilized by Wu *et al.* [226] overcame the challenge of limited data by incorporating transfer learning coupled with BO process. The approach empowered attaining quantitative structure-property relationships for thermal conductivity, which provided candidates possessing comparable thermal conductivities to those of SOTA non-composite thermo-plastics. Kim *et al.* [227] employed the genetic algorithm process that mimics the natural selection, creating over 100 novel polymers with a high glass transition temperature (T_g) > 500 K and bandgap (E_g) > 6 eV, which are suitable for dielectric materials for high-temperature capacitors. Moreover, researchers suggested that optimized GA parameters and the biased initial population with prior knowledge could significantly improve the GA scheme. Zhou *et al.* [228] demonstrated that a non-periodic and non-intuitive sequence of PE-PP copolymers, which was generated through the genetic algorithm, outperformed regular block copolymers in thermal conductivity. Atomistic molecular dynamics then performed the fitness evaluation of each

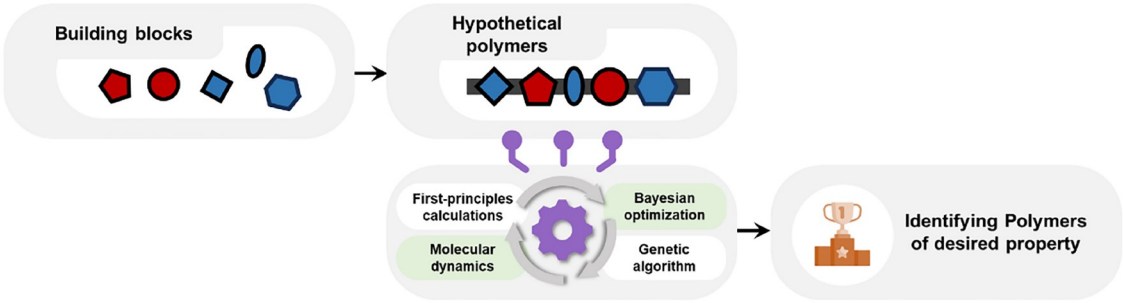

Figure 12 Schematic representation of high-throughput screening and advanced ML algorithms for target-oriented polymers (color online).

candidate by measuring its thermal conductivity.

4.3 Design and optimization of polymer synthesis

The advent of synthetic plastics in the last century revolutionized the chemical industry and the world at large. Polymer materials are now ubiquitous in our daily lives. However, traditional polymer synthesis is a process fraught with trial and error. Polymerization condition optimizing and catalyst screening are time- and resource-consuming endeavors. Moreover, this trial-and-error approach generates a significant amount of chemical waste, posing environmental concerns.

Optimizing polymerization conditions is not a simple task with a singular focus. Chemists often need to balance parameters like chemical composition, molecular weight (MW), and dispersity ($Đ$) for superior material properties. This multi-parameter, multi-objective optimization is a monumental task, compounded by the complexity of high-dimensional data, which often hinders chemists from precisely attaining diverse polymer targets. From this perspective, the potential for implementing AI and ML in polymer synthesis is enormous. By employing advanced data analysis techniques and predictive models, AI can assist scientists in rapidly identifying optimal polymerization conditions and catalysts, thereby reducing the cycle of experiments and saving time significantly.

However, the application of AI in polymer synthesis has been slower compared with its usage in optimizing small molecule organic synthesis. A primary reason is the lack of sufficient high-quality data in polymer research. The scarcity of data stems from the complexity of the polymerization process: The multi-parametric conditions and intricate polymerization mechanisms make computational simulations challenging, rendering simulated data unavailable. Additionally, stringent and sensitive polymerization conditions lead to significant variability in outcomes between different batches. These issues of data limitation pose challenges for ML modeling of polymer synthesis. Therefore, one of the key tasks for data-driven modeling of polymer synthesis relies on the acquisition of large quantities of high-quality, repeatable, and interpretable data that adhere to the FAIR (Findable, Accessible, Interoperable, and Reusable) principles [229].

One way to obtain data is searching from handbooks or literature. However, inconsistent and sometimes even contradictory results across different publications are not uncommon. High-throughput computational simulations or virtual screening is another approach. However, this approach highly depends on the computational power and is still challenging to predict experimental outcomes in a quantitative manner, especially for polymer synthesis. High-throughput experiment is a viable approach ensuring experimental consistency but relies on automation and is not suitable for experiments with long measurement times or complex material handling steps.

Flow chemistry is currently the primary choice in AI-assisted polymerization processes optimization for its ability of real-time monitoring, time-dependent data acquisition on polymer MW and monomer conversion. For example, Junkers and colleagues developed an automated flow synthesis platform for polymer synthesis, coupled with real-time monitoring using gel permeation chromatography (GPC) [230], NMR [231], and Fourier-transform infrared spectroscopy [232]. This platform enables rapid and efficient screening of reaction parameters, including residence time, monomer concentration, polymerization degree, reaction temperature, and monomer conversion rate. They used single-objective ML optimization algorithms to dynamically adjust reaction parameters, optimizing the reaction to precisely control MW or monomer conversion rate, leading to significantly reduced experimental cycles and development time. To achieve multi-objective closed-loop optimization of polymer synthesis, Warren and colleagues demonstrated an ML-assisted automated flow polymerization synthesis platform that can autonomously determine optimal polymerization reaction conditions toward predetermined polymer properties [233]. It features a computer-controlled flow reactor that autonomously polymerizes, using real-time NMR and GPC for polymer characterization. This platform utilizes

the Thompson Sampling Efficient Multi-Objective Optimization (TSEMO) algorithm to optimize reversible addition–fragmentation chain transfer (RAFT) polymerization of different monomers, exploring the trade-off between $Đ$ and monomer conversion rate. Hartman and colleagues [234] also combined automated microfluidics with ML to explore the reaction space of olefin free radical polymerization catalysts, accelerating the discovery of optimal catalytic efficiency conditions.

In analyzing the relationship between polymerization parameters and outcomes, Chen and colleagues [235] developed an ML-assisted systematic polymerization planning (SPP) platform for intelligent control of polymer MW and $Đ$. They constructed an ML model to analyze and optimize the reversible deactivation free radical polymerization process, combining multivariate analysis to uncover complex interactions between polymerization conditions for optimal polymerization condition prediction (Figure 13). Wilson and colleagues [236] also employed active learning and Bayesian optimization algorithms to accelerate the optimization of electrochemical atom transfer free radical polymerization reactions, which significantly improved the experiment efficiency.

4.4 End-to-end prediction of polymerization

Optimization of polymer conditions or catalysts for desired MW and $Đ$, however, is often not the end of polymer synthesis in real-world practices. The ultimate goal of AI-assisted polymer synthesis, as previously mentioned, is to identify targeted (multi-)functions from high dimensional, enormous chemical spaces, elucidate the hidden structure-function relationships, and accelerate material design. Thus, many laboratories in recent years have been dedicated to the development of closed-loop high-throughput "design-synthesis-test-learn" toward end-to-end AI-assisted prediction from polymer synthesis to properties/functions. This approach, again, highly relies on robust high-throughput polymer synthesis platform amenable to automation and data digitalization for successive AI/ML.

There are two primary strategies for high-throughput polymer synthesis, namely parallel copolymerization of various monomers and post-polymerization modifications. For the parallel copolymerization approach, the main challenges include: (1) polymerization reactions that are sensitive to moisture and/or air, increasing difficulties for automated liquid handling systems; (2) poor polymerization control, resulting in low repeatability and predictability; (3) limited flexibility in structural units, restricting monomer types and chemical space; (4) complex and inefficient post-treatment operations, difficult to pursue high throughput. To tackle these challenges, researchers have developed various water- and oxygen-resistant controlled free radical polymerization platforms, such as Enz-RAFT [237–240], oxygen-tolerant atom transfer radical polymerization (ATRP) [241], PET (photoinduced electron/energy transfer)-RAFT [242,243], and others [244–248], most of which overcome these issues and enable controlled high-throughput preparation of polymers. The second strategy is based on post-polymerization modification, which has been one of the primary methods for preparing high-throughput polymer libraries in recent years [249]. Efficient Huisgen cycloaddition

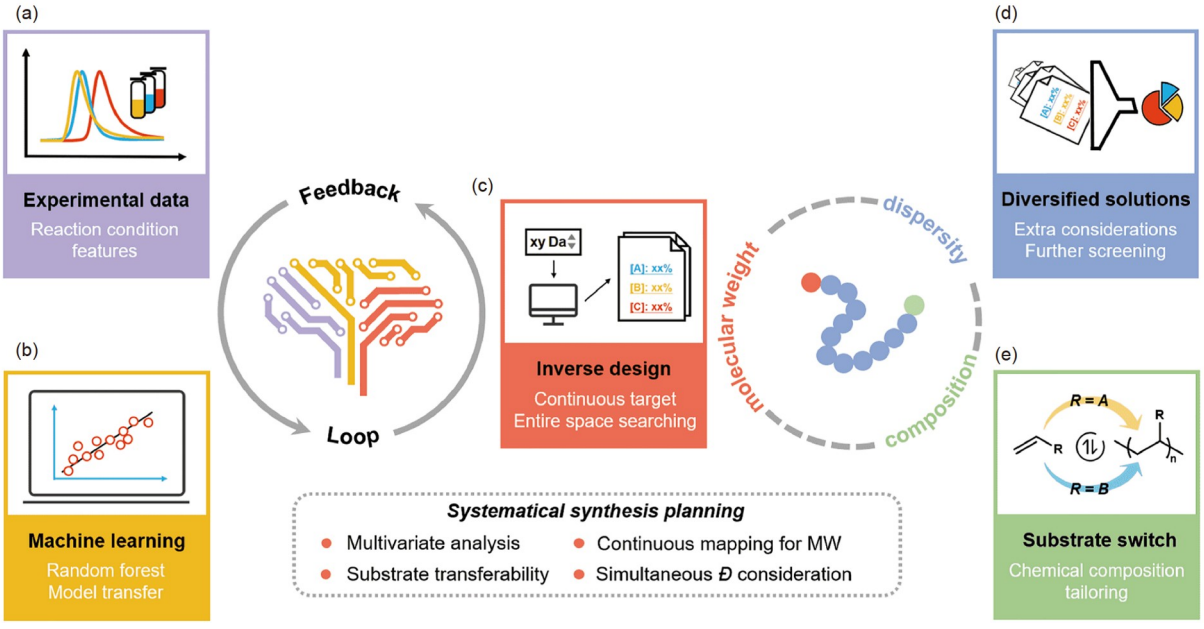

Figure 13 An ML-assisted systematical polymerization planning (SPP) platform for polymer inverse design. Reproduced with permission from Ref. [235]. Copyright 2021, Science China Press (color online).

[250], activated ester-amine coupling, thiol–ene reactions [251], and Michael addition are the most commonly used post-polymerization modification methods. The challenge with this strategy lies in the typically poor water solubility, instability, and difficulty in long-term storage of the precursor polymers.

For efficient discovery of polymers with certain functions, quick and convenient polymer purification methods are also needed in addition to high-throughput synthetic techniques. Gormley's group [252] developed a gel filtration chromatography technique that rapidly and high-throughput purifies polymers, with over 95% removal of small molecule impurities and about 85% retention rate for 32 types of polymers. However, many polymer purification strategies (such as precipitation, extraction, and chromatographic separation) often depend on specific properties of the target polymers [253,254]. And due to the complexity of these processes, the purification step was sometimes skipped in some cases.

Once a vast amount of structural and informational data were successfully generated through the high-throughput synthesis and characterization methods, the next key issue is how to effectively mine these data to guide new material design. Applying ML to identify key features from past data can guide studies on material structure-function relationships [255,256]. For instance, Reineke and colleagues [43,257] combined polymer design with parallel experimental workflows to discover efficient polymers for intracellular ribonucleoprotein (RNP) delivery. Utilizing interpretable ML, they computed SHAP (Shapley additive explanations) for nine polymorphic features, uncovering the structure-function relationship behind editing efficiency, cytotoxicity, and RNP uptake, providing guidelines for designing polymer libraries based on RNP delivery. Bao *et al*. [258] reported an ML-assisted method that guides the design of full-color tunable emission trans-space charge transfer through-space charge transfer (TSCT) polymers. They synthesized 71 different chain length and type styrene polymers through ATRP, building Maximum Likelihood Expectation Multivariate Linear Regression (MLREM) and Bayesian Regularized Artificial Neural Network (BRANNLP) models to predict the photophysical properties of unknown TSCT polymers, exploring the relationship between structure and function. Olsen *et al*. [259] used high-throughput synthesis techniques to create a large library of 642 polyesters and polycarbonates, while developing a high-throughput clean area biodegradation test to assess the biodegradability of the polymers. They used ML models to interpret the structure-property relationships of polymer biodegradability. Knight and colleagues [260] designed and synthesized a series of polymers containing novel triphenylphosphine acrylamide monomers, using ML regression models to study the relationship between polymer properties and polymerization catalysis rates.

In recent years, the ML-assisted closed-loop HTE has shown immense potential in the discovery of new materials [261–263] (Figure 14a). For example, Leibfarth and colleagues [264] combined ML with flow polymerization, enhancing the magnetic resonance signal strength of fluorinated polymers through only about 300 experiments and discovering previously unreported structure-effect relationships (Figure 14b). Gormley *et al*. [265], through a closed-loop high-throughput polymerization and active learning strategy, rapidly discovered polymers for neuroregeneration research that could protect proteins, as well as designed stable proteinase-active random copolymers [266] (Figure 14c). Lu

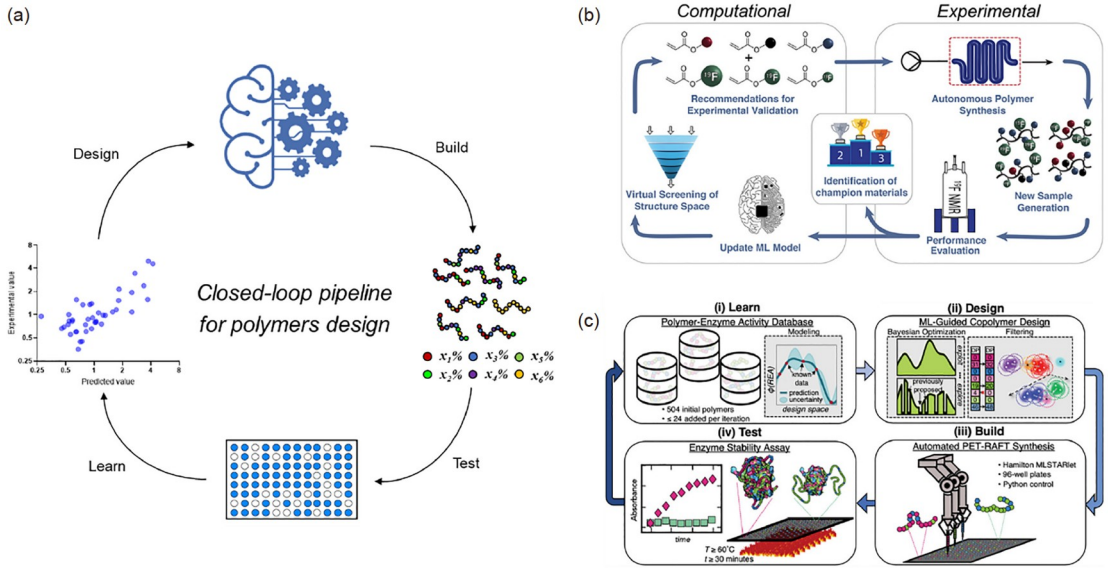

Figure 14 (a) An ML-assisted design-build-test-learn closed-loop pipeline for the evolution of polymers. (b) Active-learning-guided discovery of copolymer ^{19}F MRI agents. Reproduced with permission from Ref. [264]. Copyright 2021, American Chemical Society. (c) Closed-loop design-build-test-learn process for the design of polymer–protein hybrids. Reproduced with permission from Ref. [266]. Copyright 2022, John Wiley & Sons (color online).

and colleagues [267] established a high-throughput post-polymerization modification platform for selenium-containing polypeptides synthesis. By incorporating ML algorithms, they were able to efficiently explore the functional chemical space of 600 random copolymers for desired functions such as enzyme-like catalysis without much prior knowledge in four days (Figure 15). It is foreseeable that the deep integration of high-throughput technology and ML will have a significant impact on polymers and will aid in accelerating the discovery of materials in key areas.

4.5 AI Application in biological macromolecules

As mentioned previously, synthetic polymers are highly heterogeneous whose structural information is hard to encode. By contrast, biological macromolecules (such as nucleic acids and proteins) are highly programmable, which provides an exciting arena for AI application. All the details about their structures and functions are conveniently encoded in sequences and facilely manipulated by a set of biochemical tools. The structure-property relationship can thus be deduced from the mapping between sequence and function. Consequently, molecular engineering of biopolymers is usually accomplished with precise sequence variation. The astronomically large sequence space inevitably brings in unparalleled complexity and "the curse of dimensionality" in research and engineering. In view of the enormous biological data accumulated over the past decades (*e.g.*, PDB, UniProt, UniClust, BFD), it is an ideal scenario for the use of AI. To date, the application of AI has already transformed many fields of bio-macromolecular research, particularly in protein sciences such as protein structure prediction, de novo protein design, and protein engineering.

Proteins perform functions through their native structures. The Anfinsen's dogma postulates that the native structure of one protein is determined only by the amino acid sequence as the thermodynamically most stable structure. The structure prediction thus composes the "protein folding problem". Classically, this is accomplished by developing a reliable energy function and efficient conformational sampling protocol, as exemplified by the Rosetta software. The advent of AI-based methods pushed the structural modeling quality to approach that of experimental accuracy, resulting in a 1000-fold increase in structural data [268]. AlphaFold2 [9] and RoseTTAFold [269] learn evolutionary information from multiple sequence alignments. The use of protein language models such as ESMFold overcomes the limitation to generalize across protein families and facilitated atomic level prediction from single sequences [270]. While certain limitations remain, such as overpresentation of proteins in spite of missing features (cofactors, post-translational modifications, partners), insensitiveness to mutations, and incapability of generating dynamic ensembles, protein structure prediction seems a largely solved challenge. Predicting function from sequence using ML has also been demonstrated by assigning the enzyme commission (EC) number for a given sequence [271]. The availability of more structures further allows genome mining based on structures using tools like Foldseek [272]. These tools greatly expand our knowledge about proteins.

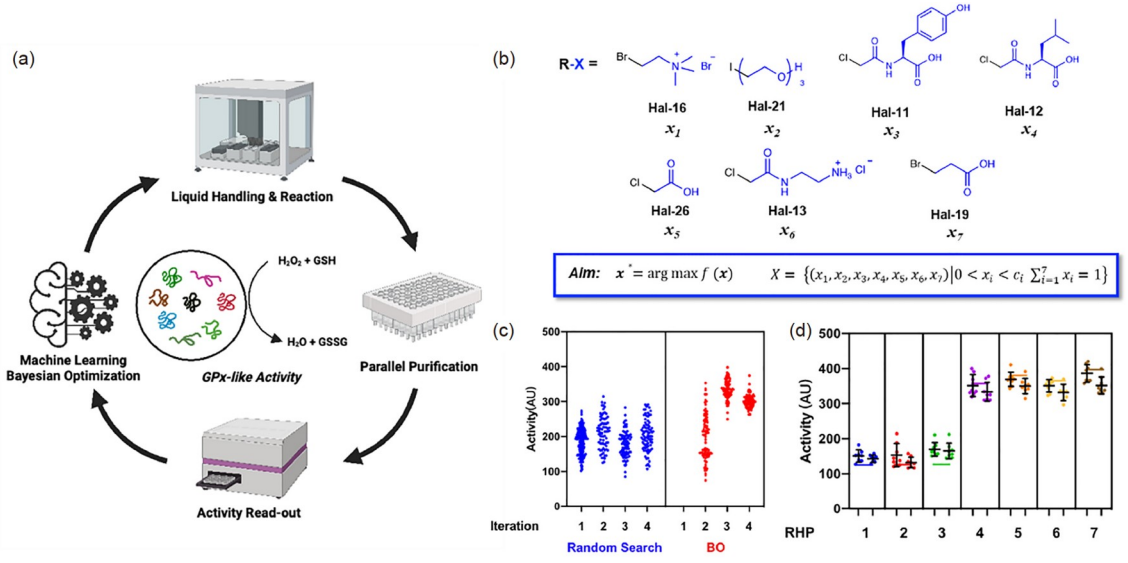

Figure 15 (a) Closed-loop optimization of GPx activity of the heteropolypeptides *via* high throughput synthesis and machine learning. (b) Structure of the seven selected organohalides for heteropolypeptides library generation and aim of optimization. (c) GPx-like activity of RHPs in each iteration *via* random searching (blue) or Bayesian optimization (red). (d) Data validation within a plate (n = 8) and between two different plates. RHPs with low (lanes 1–3) and high (lanes 4–7) GPx-like activities from the database were selected for validation. The dots on the right and left side in each lane represent the results from different plates. The black central lines and error bars in each lane represent the mean and s.d. The coloured line in each lane is the original activity of the RHP from the database.Reproduced with permission from Ref. [267]. Copyright 2023, Nature Publishing Group (color online).

The "inverse folding problem" of protein design aims at finding amino acid sequences that fold selectively into a desired "target" structure. More broadly, *de novo* protein design focuses on generating structures new to our knowledge or accomplishing functions (*e.g.*, binding, fluorescence, catalysis) new to the scaffold. Currently, there are mainly two approaches to design, namely, data-driven and physics-inspired. The former relies on sequence features that can be extracted and leveraged by various neural network-based generative models, such as UniRep [273], ProGen [274], ESM-1b [275], ProViz [276], ProtTrans [277], and ProteinBERT [278], for structure or sequence generation. The latter combines free energy calculation, binding affinity calculation, or conformational entropy estimation with sequence variation for de novo design toward a target structure/function. It includes force-field-based methods like Rosetta [279] and FoldX [280] and ML-based tools like ABACUS [281] and ProteinMPNN [282]. To evaluate the designability of a protein fold, a backbone centered energy function of neural network, SCUBA [283], was developed. When the target structure is partially/fully absent, hallucination protocols can be employed based on trDesign [284], RFdiffusion [285], and Chroma [286]. In addition, it is also possible to perform controllable generation of proteins directly at the sequence level. The convergence of these two complementary approaches has proven synergistic and powerful in achieving hard goals such as *de novo* enzyme design [287,288]. The problem of these methods is the relatively low success rate, which mandates labor-intensive screening of hundreds to thousands of designs for validation. This challenge may be ameliorated by high-throughput robotic automation. While some of the obtained structures can agree precisely with the design at the atomic level, it is often difficult to gain the desired function with high activity as designed, which necessitates further rounds of directed evolution.

Directed evolution comprises two steps: library generation and property screening. Traditionally, directed evolution takes an uphill hike on the protein fitness landscape by accumulating beneficial mutations over rounds of mutation/screening. With high-throughput sequencing techniques and low-cost assay methods, the information about otherwise discarded suboptimal mutants can also be used to train ML models to capture the sequence-function relationship. When meaningful features are included in the representations, simple ML models such as linear regression or shallow neural network could work well, especially for those with highly correlated local mutations [289–292]. State-of-the-art protein language models can be pre-trained on sequences from all protein families and fine-tuned with multiple sequence alignments of homologues so as to be more task-specific [275]. Notably, the limited number of experimental assays (often <100) presents a considerable challenge for high-accuracy prediction using ML models. To cope with the "low-N" scenario or even enable high-accuracy zero-shot predictions, one could combine assay-labeled data and ML models trained under different contexts (*e.g.*, probabilistic context, evolutionary context, structural-aware context) [293]. Such fitness predictors can navigate through the enormous fitness landscape by strategic virtual screening or by steered generative models [294]. To ensure broad landscape coverage with carefully chosen premium designs, one can use either a straightforward greedy algorithm or Beam search or Bayesian optimization to gain an acceptable trade-off between exploitation and exploration. It is not surprising to see that AI has already contributed considerably to directed evolution. But, protein complexity still demands heavy wet-lab experiments for directed evolution.

To reduce the workload of directed evolution, continuous evolution schemes such as PACE have been developed [295], which leads to the discovery of powerful molecular biology tools like RNA polymerases [296] and base editors [297]. However, as a platform, it relies on the cell survival for selection and can hardly be adapted to non-living circumstances. To meet the enormously diverse need of protein engineering, it is increasingly recognized that an integrated biofoundry platform combining core robotic instruments (like liquid handlers, thermocyclers, fragment analyzer, and colony pickers) and AI algorithms (for data analysis and decision making) would be indispensable to enable a closed-loop *in vitro* continuous evolution [298] (Figure 16). The biochemical processes for making biological macromolecules are usually robust under mild conditions, which is ideal for implementing automation. For example, PlasmidMaker has been developed as an end-to-end pipeline for automated plasmid construction [299]; BioAutomata has been developed as a closed-loop system for microbial pathway engineering [300]. Although it remains nontrivial to adapt biofoundary to diverse assay methods, this system is highly promising in terms of high-quality data generation, acquisition, and analysis. They can be used to quickly evolve the ML model to be more and more powerful over time. The interactive interface can be in the form of an AI agent specialized in protein sciences which conveniently communicates with personnel in human language. Eventually, a paradigm shift in protein engineering is envisioned. It is only with Biofoundry that the need for speed in industry could be potentially met. Ready access to diverse enzymes shall bring yet another revolution to the synthesis of small molecules with time frame and cost superior to chemical methods.

The above mainly focuses on protein as the model biological macromolecule. This is also where most literature works on. In principle, the work could be similarly done on RNA molecules. There are also works on using ML for structure prediction of RNA molecules [301–303]. Polysaccharides are an exception since they are not genetically encoded and highly heterogeneous. Hence, they behave more

like synthetic polymers discussed in the previous section with collective materials properties as a functional output. Nevertheless, their chemical structures may be precisely manipulated by various enzymes like glycosynthase and glycosyl transferases, and their precise structures have great implications in cell signaling and are heavily involved in diverse biological pathways. Their synthesis and structural editing may be performed using biofoundary in a way similar to proteins [304]. Overall, AI has impacted and will continue to influence the engineering of biological macromolecules for diverse purposes, especially when aided with closed-loop automation (Figure 16).

5 Automated experimentation

For AI applications in synthetic chemistry, the generation of large-scale, high-quality chemical synthesis data is not only a crucial foundation for chemical modeling but also a vital knowledge source driving the innovation of synthetic chemistry itself. However, the approach of generating synthetic chemistry data has not undergone revolutionary changes over the past century. In current chemical experimentation, manual operations still dominate, which not only makes the synthetic exploration labor-intensive but also limits the efficiency of experiments and the reproducibility of synthesis data. The advent of autonomous synthesis platforms offers a novel strategy to address these issues. These platforms, by integrating advanced control technologies and robotic systems, are capable of precision control over the chemical synthesis process, thereby enhancing the efficiency of synthetic experiments, reducing labor input, and ensuring the accuracy and reproducibility of experimental results. Recent years have witnessed significant advancements in automated synthesis, separation, and even entire intelligent synthesis systems, providing a critical hardware engine for the paradigm shift in synthetic chemistry (Figure 17).

5.1 Automated synthesis

Automation provides an avenue to transfer organic synthesis from a labor-intensive job to a machine-driven process [305]. The first concept of automated synthesis can be traced back to the 1960s when Merrifield and Stewart reported an automated system for solid-phase peptide synthesis [306,307]. Taking advantage of a similar strategy, DNA [308,309], RNA fragments [310], as well as polysaccharides [311] could be synthesized through an automated procedure today.

For peptides and oligonucleotides, the synthetic protocol

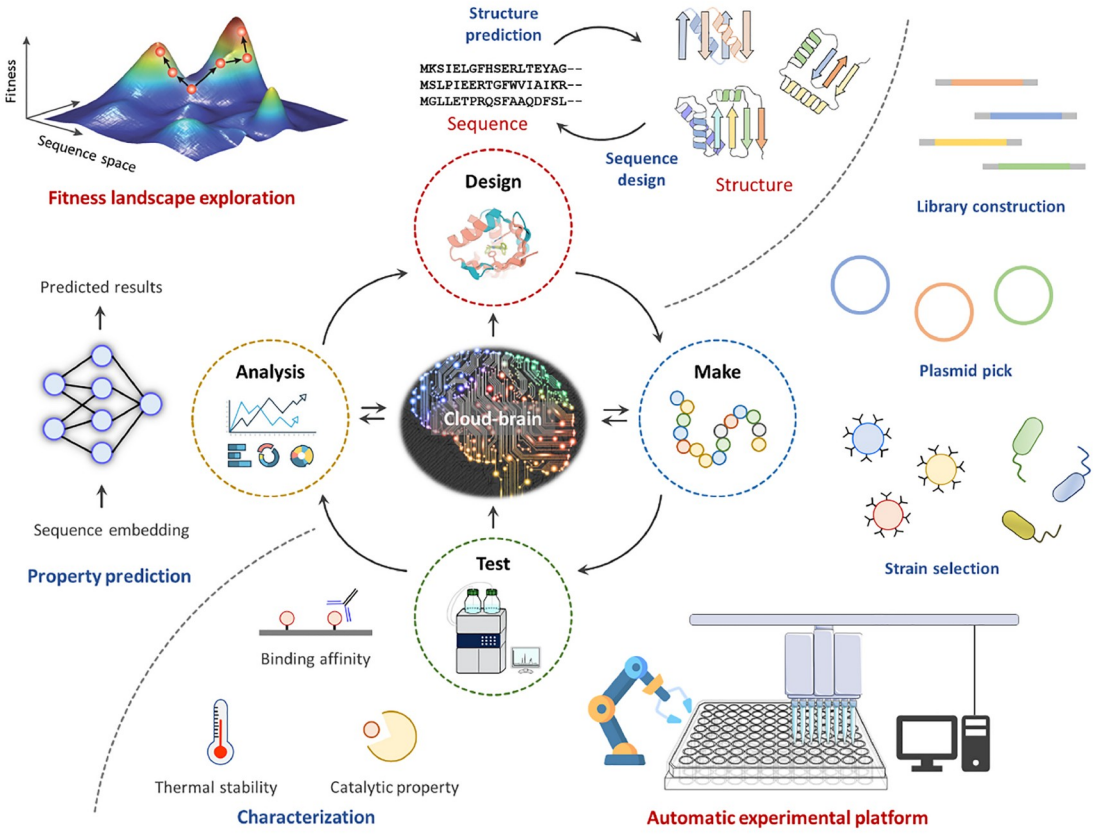

Figure 16 Closed-loop, *in vitro* continuous directed evolution enabled by AI-assisted protein design and robotic automation (color online).

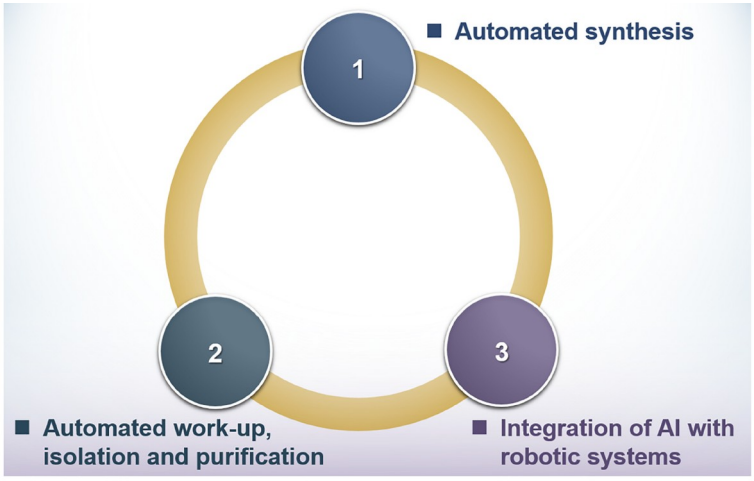

Figure 17 Key research directions of automated experimentation (color online).

in their automated synthesizers is fundamentally the same for every individual molecule. In contrast, this is not the same case for general molecular synthesis. The synthesis instruments need to adjust the synthetic routes to holistic, interdependent, and multistep processes, which are mostly distinctive for each synthesis of small organic molecules. Given that automation in chemical research is rare and the commercially available systems were usually designed for specific purposes and only valid for repetitive work. In 1978, Legrand and Foucard [312] developed an automation kit for synthetic chemists. Ley's group [313] devised a convenient and efficient prototype for evaporating, concentrating, and switching solvents in continuous flow processes and batch mode. In 2013, researchers from AbbVie Inc. developed an efficient compound-synthesis system with integrated components and automated sample-handling modules [314]. Tu *et al*. [315] later developed a fully automated synthesis-purification station based on the SWAVE platform and in-house developed robotics. Very recently, Ahmed's group [316] developed a robot-assisted acoustofluidic end effector (RAEE) system consisting of a robotic arm and an acoustofluidic end effector (Figure 18a).

Continuous flow manufacture is widely embraced for synthesizing active pharmaceutical ingredients (APIs) and fine chemicals (Figure 18b) [317,318]. Automated platforms, such as ChemKonzert (Figure 18c) [319], enable solution-phase synthesis of diverse organic compounds. Pentelute's lab introduced an automated flow-based system for rapid polypeptide synthesis [320]. Pfizer's platform integrates nanomole-scale screening and micromole-scale synthesis, conducting over 1,500 experiments per 24 h [321]. Li *et al*. [322,323] developed Tiny Tides, a fully automated fast-flow device, achieving on-demand customized antisense phosphorodiamidate morpholino oligomers (PMOs) and high-speed synthesis of PPNAs. This high-efficiency synthesizer serves as a training data source for effective ML models guiding efficient PNA sequence design [324].

In 2019, Cronin *et al*. [325] developed an autonomous compiler and robotic laboratory platform, called Chemputer (Figure 18d), to synthesize organic compounds on the basis of standardized methods descriptions. Dömling's group [326] employed I-DOT, a positive-pressure-based low-volume dispensing technology, for fully automated synthesis of over 1,000 iminopyrrolidine-2-carboxylic acid derivatives through Ugi-3-component reaction at the nanoscale. Williams, Kappe, and colleagues [327] designed an integrated multistep reaction and real-time analysis platform for controlled synthesis of mesalazine, achieving a throughput of 1.6 g per hour. In 2021, Kim's group [328] developed a parallel flow synthesizer enabling multiplex synthesis and optimization of compound libraries, offering rapid screening and obtaining optimal conditions for various reactions in less than one hour from 96 different conditions. Gilmore and colleagues [329] introduced an automatic radial synthesizer featuring multiple continuous flow modules arranged around a central core, enabling stable and reproducible linear and convergent syntheses without manual reconfiguration. In 2022, Jensen *et al*. [330] developed a continuous stirred-tank reactor (CSTR) flow platform capable of handling solids and slurries during chemical transformations, enhancing the identification of optimized reaction conditions for manufacturing process development.

Isolation and purification in flow chemistry can follow an ideal process where reactants enter, and pure products exit continuously. George *et al*. [331] demonstrated continuous artemisinin synthesis in a supercritical CO_2 flow system. Multi-step reactions often require interruptions for work-ups and extractions before proceeding. Inline solid-phase extraction [332], gas-liquid, and liquid-liquid separation [333] technologies can incorporate most work-ups into a con-

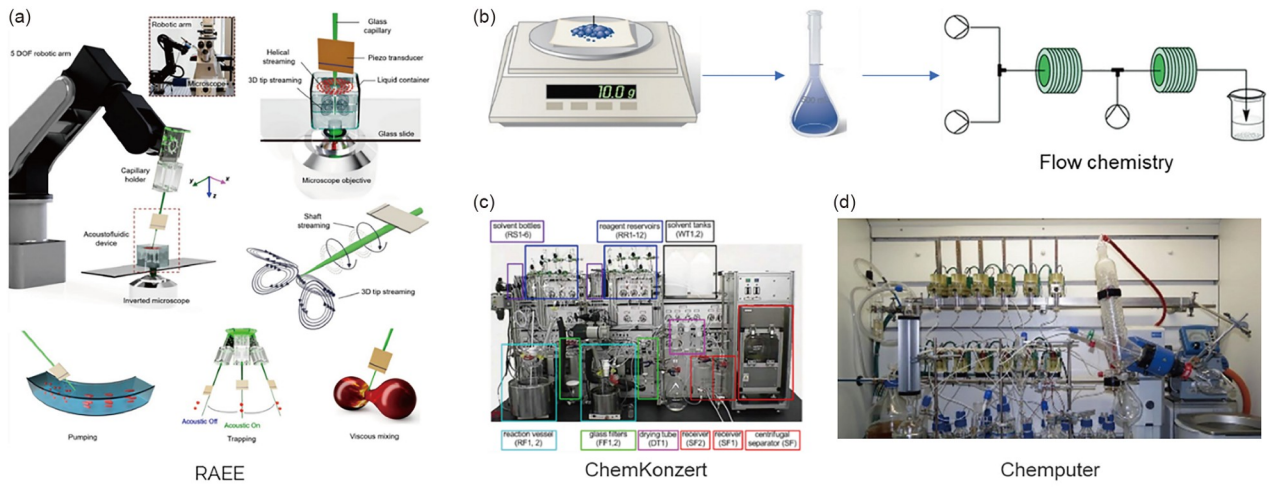

Figure 18 Schematics of flow chemistry-based automatic synthesis platforms. (a) RAEE system. Reproduced with permission from Ref. [316]. Copyright 2022, Nature Publishing Group. (b) Flow manufacturing. Reproduced with permission from Ref. [317]. Copyright 2017, American Chemical Society. (c) ChemKonzert system. Reproduced with permission from Ref. [319]. Copyright 2010, Pharmaceutical Sociey of Japan. (d) Chemputer system. Reproduced with permission from Ref. [325]. Copyright 2019, American Association for the Advancement of Science (color online).

tinuous process. Baranczak *et al*. [334] developed a fully automated platform for synthesis-purification-testing of small molecule libraries. Lee and Vilela *et al*. [335] reported an inline chromatographic purification automated flow synthesis platform, achieving 97%–99% purity in continuously isolating products.

Burke's automated Lego-like synthesis process utilized iterative peptide coupling for Suzuki-Miyaura $C(sp^2)$–$C(sp^2)$ bond formation [336], creating 14 diverse small molecule classes. The approach used *N*-methyliminodiacetic acid (MIDA) as a building block, employing a "catch-and-release" purification protocol [337]. The strategy, while incompatible with stereospecific $C(sp^3)$–$C(sp^2)$ or $C(sp^3)$–$C(sp^3)$ bond-forming reactions due to MIDA sensitivity, was recently improved with stable tetramethyl-*N*-methyliminodiacetic acid (TIDA) boronates [338]. This advancement enabled the automated synthesis of $C(sp^3)$ boronate building blocks and facilitated stereospecific $C(sp^3)$–C bond formation, broadening the scope of accessible molecules [339]. Jensen and collaborators [340] pioneered microfluidic automated platforms, such as droplet-based systems for efficient reaction screening and product isolation in small-scale medicinal chemistry. They optimized Pd-catalyzed C–N coupling conditions [341] and developed an automated single-droplet screening platform for electroorganic process discovery [342,343]. In 2020, Kennedy and Stephenson *et al*. [344] reported an automated microfluidic platform to enable picomole scale synthesis. Recently, Jensen and Pidko *et al*. [345] reported a catalytic asymmetric hydrogenation of a sensitive β-amino-ketone substrate by means of an automated microfluidic platform. Debrouwer *et al*. [346] reported a dual catalysis cross-electrophile coupling using oscillatory plug flow photoreactors.

Adamo *et al*. [347] introduced a compact continuous manufacturing platform. Bode's group [348] developed an automated capsule-based synthesis for *N*-heterocycles. Bode *et al*. [349] designed an iterative console assembling molecules from vast virtual libraries. Cronin *et al*. [350–352] utilized 3D printing for interconnected modules and a chemical to computer-automated design (ChemCAD) approach. They later created a portable platform for universal chemical synthesis using chemical markup language (χDL) and 3D printing. These advancements signify a transformative shift towards efficient, digitized, and automated synthetic platforms [353].

Various research groups have explored the concept of a "cloud lab" for remote operation of self-optimizing systems. In this regard, Poliakoff's group [354] demonstrated a remote-operated system. Ley's group [355] introduced Ley Lab in 2016, an Internet-based software allowing global monitoring and control of chemical reactions. Aspuru-Guzik and colleagues [356] developed ChemOS in 2018, a portable framework employing AI, sensors, and robotics for closed-loop systems. Cronin's Chemputer translated reported procedures into automatable steps using NLP and χDL [325,357]. Zhu's materials acceleration operation system (MAOS) [358] in 2020 enabled intelligent robotics for material synthesis with AI-controlled quality assurance, accessible through VR-robot interaction. Cooper's 2020 robo-chemist [359], driven by a Bayesian algorithm, autonomously conducted 688 reactions over eight days. Jiang's AI-Chemist can autonomously extract literature and propose experimental plans from a cloud database [360]. In 2023, Gomes *et al*. [7] developed a system called Coscientist which is an artificial intelligence system driven by GPT-4 that autonomously designs, plans, and performs complex experiments.

5.2 Automated work-up, isolation and purification

Automated work-up, separation, and purification platforms are integral components of laboratory automation. In this regard, Ley *et al*. [361] has done significant contributions, whose works have been comprehensively reviewed by their own review. For instance, they have implemented machine vision automation for extraction operations [362], online solvent flash evaporation devices [313], optimized chromatographic separations [363], and automated filtration [364]. In terms of automating chemical laboratories and integrating ML and deep learning, the Cronin research group has achieved remarkable progress for automated automated synthesis machine [325,365]. Their system's implementation relies on computer-controlled pumps. These pumps inject reactants into reaction flasks. Reaction work-up, including extraction, column chromatography, and rotary evaporation, is also integral to the system. These operations are achieved by transferring liquid reactants through a complex pipeline system using pumps. Spectroscopic detection methods, such as infrared spectroscopy and nuclear magnetic resonance spectroscopy, are also integrated. ML algorithms are employed to interpret these spectra, obtaining reaction information, which is then fed back into the system to achieve a closed-loop optimization.

In chromatographic analysis and preparation, Kassel *et al*.'s PrepLCMS [366], a pioneering mass spectrometry-based system, automates the purification of substantial compound quantities. Koppitz *et al*.'s LC/MS-based system efficiently processes 100–200 compounds daily [367], ensuring high purity and yield. Ilg *et al*. [368] introduced a high-throughput high performance liquid chromatography/mass spectrometry (HPLC/MS) platform, incorporating Covaris technology for sample preparation, automated aliquotation in fractionation, and a novel evaporation technique combining freeze-drying, enhancing purification efficiency. Recently, Mo *et al*. [369] developed an automated thin layer chromatography (TLC) platform for high-throughput data collection, subsequently using ML methods to predict the retardation factor (Rf) of compounds. The trained ML model can accurately predict the Rf value curves of organic compounds under different solvent combinations, providing general guidance for purification condition selection. Additionally, they have also developed a QGeoGNN-based model for predicting optimal HPLC separation conditions for chiral enantiomers, significantly reducing trial-and-error costs [370].

5.3 Integration of AI with robotic systems

Discovering new reactions is unpredictable and laborious. Suboptimal initial conditions, especially in micro/nanoscale, may lead to overlooked trace products. In this regard, the integration of AI with robotic system can provide an effective strategy. However, it should be noted that applying ML to navigate new chemical space is underexplored due to the challenge of assessing reactivity in unknown reactions with unpredictable products compared with optimizing conditions for known target compounds [371,372]. Deconvolution algorithms, for instance, can help identify novel products [373,374]. Cronin's group [365] demonstrated that a synthetic robot controlled by SVM algorithm significantly accelerated organic reaction discovery. The liquid-handling robot selected reactants from a pool, with real-time analytics monitoring reactions. ML built a chemical space model, recommending experiments and controlling the robot. The system outperformed manual processes, predicting the reactivity of 1,000 combinations with over 80% accuracy. Zahrt and colleagues [375] applied ML to guide electrochemical reaction discovery, developing a molecular representation for general models and successfully predicting new reactions' competency. These studies showcase AI-driven chemical robots advancing reaction space exploration. Recently, research groups have also applied reinforcement learning for automated mechanism discovery, bypassing exhaustive screening [376,377]. An agent constructs efficient reaction pathways by selecting actions (elementary steps) with varying rewards. This approach holds promise for efficient reaction network exploration, requiring first-principles or semi-empirical evaluations.

6 Challenges and perspective

The burgeoning field of AI in organic and polymer synthesis presents a transformative potential for scientific discovery. However, this promise is contingent on overcoming a series of challenges that currently impede its full realization. This section delves into these critical issues, offering a succinct yet comprehensive overview of the challenges faced by AI applications in synthetic chemistry as well as potential solutions.

6.1 Data

In synthetic chemistry, particularly in organic and polymer synthesis, the role of data is foundational for the successful AI application [378]. The main challenges associated with data in this field include issues of quantity, quality, standardization, and accessibility [229,379,380]. Generating sufficient, high-quality data is a complex endeavor, limited by the intricate and time-consuming nature of chemical experiments. The quality of data, essential for the training and performance of AI models, is frequently compromised by variations in experimental conditions, disparate practices among researchers, and inherent biases, leading to sig-

nificant inconsistencies [381–383]. These variations and the lack of detailed reaction conditions in public databases undermine the reliability of data for AI applications. Moreover, the absence of standardized data formats complicates the compatibility and comparability across different systems, hindering the efficient training of AI models and their application to varied tasks. Data accessibility is further challenged by legal, technical, and proprietary barriers that restrict the use of data, making it difficult for researchers to obtain and utilize the information needed for their work.

Addressing the limitations around data in synthetic chemistry necessitates a multifaceted approach that integrates the establishment of open data principles with advanced AI-assisted data management techniques. The adoption of FAIR principles—ensuring data is **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable—is critical for improving data quality, standardization, and accessibility [384]. These principles support the creation of a standardized data management framework that facilitates the sharing and reuse of data across the scientific community. Additionally, leveraging AI for automated data extraction and processing offers a powerful solution to enhance the efficiency and accuracy of data collection [385,386]. This involves the use of advanced natural language processing and large language models for scraping, mining, and extracting valuable information from a plethora of sources including chemical literature, reaction databases, and experimental records. The key to harnessing these technologies lies in their ability to process and analyze vast amounts of data, translating them into actionable insights that can drive research forward. However, ensuring the reliability of the extracted data is crucial, necessitating careful validation and verification processes. Moreover, fostering an open data community, grounded in the principles of collaboration and shared resources, is essential for overcoming the barriers of data accessibility and standardization. Such a community would serve as a hub for aggregating, refining, and sharing data, thereby facilitating a more collaborative, efficient, and innovative research environment [35]. Together, these strategies offer a comprehensive blueprint for addressing the challenges posed by data limitations in synthetic chemistry, paving the way for enhanced AI applications and scientific discovery.

6.2 Encoding

For the digital representation of synthetic chemistry, three core challenges are prominently identified: universality, interpretability, and the representation of the stochastic nature of polymer structures. The complexity and diversity of chemical data, spanning a wide spectrum from molecular structures to reaction conditions, necessitate distinct representation approaches for each type, complicating the quest for universality. The issue of diversity, encompassing various modalities such as texts, images, and tables, adds another layer of complexity in standardizing chemical information. Further complicating this pursuit is the fact that various laboratories and researchers often employ their customized methods for data recording and representation. These personalized approaches create significant hurdles in achieving a universal standard for chemical data across different formats and sources. The challenge of interpretability arises from the need to encode chemical insights in a way that is comprehensive to computational models and intelligible to human researchers. This includes difficulties in conveying complex chemical phenomena and the inherent tension in designing models that combine high accuracy with ease of understanding, emphasizing the trade-off where increased predictive performance often diminishes transparency. Additionally, accurately capturing the stochastic nature of polymers, characterized by their varied molecular weights and structural configurations, presents a unique challenge. The properties of polymers are heavily influenced by their molecular diversity, requiring nuanced and precise encoding strategies to capture the essential characteristics that dictate their behavior and functionality. These challenges collectively underscore the complexities of developing effective encoding systems in synthetic chemistry, aimed at bridging the gap between the intricate chemical phenomena and their computational representations.

Advancing encoding techniques in synthetic chemistry can be approached from the following angles. Leveraging multimodal learning methods and large Transformer-based models such as ChemBERTa [387], MoLFormer [388], and ChemGPT [389] to integrate chemical data from diverse modalities including texts, images, and tables, could pave the way towards a unified representation system. This effort may also involve standardizing chemical information through the creation of universal datasets and the application of intelligent algorithms to address the challenges of non-standardization. On the interpretability front, enhancing models to combine high accuracy with ease of understanding is crucial. A representative example is the ASO descriptor designed by Denmark *et al.* [187], which finely depicts the three-dimensional structure of chiral molecules from the perspective of space filling. Additionally, symbolic regression represents a valuable method that could uncover relationships with clear analytical expressions, offering new ways to interpret complex chemical data [390]. For the representation of the stochastic nature of polymers, it requires models that can encapsulate the diversity in molecular weights and structural configurations. Specialized polymer representation models that consider dispersity and monomer sequence arrangements could more precisely predict polymer properties [214]. Exploring computational models for topological structures, such as branched polymers, might

also improve the accuracy of property predictions and expand the models' applicability. Through these strategies, the goal is to effectively bridge the gap between the complexity of chemical phenomena and their digital representation, facilitating AI applications for chemical understanding and innovation.

6.3 Model availability

In synthetic chemistry, the field faces the challenge of model availability due to its diverse chemical dimensions and highly individualized application scenarios. Related AI researches often narrow the focus to specific synthetic targets, utilizing customized datasets of limited size. This specialized approach to developing and implementing AI models in synthetic chemistry is not yet fully mature. Although certain AI models demonstrate significant potential, the majority presented in research papers typically provide only a GitHub link with minimal annotations. This mode of sharing, while enabling the replication of research, lacks in offering user-friendly software, platforms, or sufficient documentation and user guides, rendering it difficult for chemists without computer science expertise to effectively utilize these models. Additionally, most model developments prioritize the verification of scientific hypotheses over the consideration of the models' applicability from the users' perspective. Even successful model implementations may not meet the specific needs of synthetic chemists for particular molecules or reactions. The efficiency and accessibility of the encoding process also pose notable challenges. Many of the current models require the use of specialized quantum chemistry software and significant computational resources, further complicating matters for experimental synthetic chemists. Thus, making AI technology conveniently and efficiently usable for experimental chemists is essential for the progress of AI in synthetic chemistry.

To tackle the issue of availability, democratizing AI becomes a critical step towards technological advancement, essential for fostering scientific innovation of AI-assisted synthetic advancement. This democratization process aims to make AI tools, algorithms, and software more accessible and user-friendly for synthetic chemists, particularly those without a background in computer science. Developing AI software that aligns with the needs and experiences of chemists, moving away from complex code repositories to tools characterized by intuitive data input, clear result displays, and simple operation procedures, can significantly lower the barriers to AI application, thereby improving its impact in synthetic chemistry. Moreover, the transparency and chemical interpretability of AI tools are crucial; they should not only provide accurate predictions but also clearly explain their decision-making processes to users, building trust and promoting positive interactions between chemists and AI. In addition, allowing chemists to contribute to the AI modeling processes can lead to more meaningful predictions and ensure that AI-assisted experimental designs are closely aligned with real-world scenarios. The democratization of AI in synthetic chemistry is not just about making AI more accessible; it is about creating a more collaborative and innovative environment where AI and synthetic chemistry complement each other.

6.4 Automated experimentation

Automated or semi-automated platforms, utilizing robotics and data-driven algorithms, present a solution to the bottleneck in chemical synthesis. While automation is well-established in routine tasks for pharmaceuticals, it often focuses on narrow, well-defined processes. Methodologies for chemical synthesis automation, optimization, and discovery, particularly in laboratory-based research and bench-scale synthesis, face challenges from both hardware and software, as well as the high cost. For the hardware foundation, a critical issue is the integration of automated synthetic platforms into existing laboratory setups. This requires not only consideration of the physical space within synthesis labs but also the adaptability of the platform to seamlessly fit these environments. Moreover, the importance of user-friendly software interfaces and application programming interfaces (APIs) cannot be overstated. Chemists are in search of comprehensive solutions that encompass reaction monitoring, machine self-optimization, and AI/ML algorithms, all compatible with remote control capabilities. Currently, the software and APIs available fall short of supporting a fully self-driving laboratory, indicating a significant gap that needs to be bridged.

In tackling the challenges faced by automated chemical synthesis, a focused approach on both hardware and software innovations is pivotal. For hardware, the introduction of customizable, modular systems like Opentrons' laboratory robots for liquid handling showcases a significant step towards affordability and adaptability in automation. Their open-source robots (OT-1 and OT-2), priced as low as $10,000, exemplify the move towards making sophisticated automated platforms more accessible to a broader audience, ensuring easy integration into existing lab setups without extensive modifications. On the software side, beyond democratizing AI, the incorporation of integrated AI management software holds the key to bridging the gap in automated chemical synthesis. Systems like those developed by Jensen and Jamison, utilizing MATLAB and LabVIEW, offer examples of how control systems can provide real-time monitoring and automated feedback optimization [391]. Such platforms demonstrate the potential for AI-based synthesis planning and ML algorithms to revolutionize synthesis routes, from hypothesis generation to molecule structure

prediction. Together, these hardware and software advancements present a coherent strategy to overcome existing obstacles in automated chemical synthesis. By aligning the cost-effective, customizable hardware solutions with cutting-edge, integrated software platforms, the field is set to undergo a transformative shift towards more accessible, efficient, and innovative research methodologies, marking a significant leap in the application of automation technology within the chemical sciences.

7 Conclusions and outlook

In summary, this review discusses the applications of AI in organic and polymer synthesis in recent years, investigating the benefits and potential of the data-driven research paradigm in addressing challenges of synthetic chemistry. In organic synthesis, AI applications have made significant breakthroughs at various levels ranging from molecules to reactions: (1) Predictions of molecular thermodynamic and kinetic properties have seen a quantum leap in efficiency without sacrificing accuracy. Chemists, empowered by ML models, can now swiftly and precisely assess crucial physicochemical parameters like pK_a, BDE, and rate constants, offering valuable insights for molecular design in synthetic chemistry. (2) The capabilities of computer-assisted synthetic planning have undergone tremendous improvement, particularly for complex molecules. Emerging AI software can more rationally, diversely, and efficiently plan multi-step synthetic routes, even rivaling human chemist designs. (3) Data-driven prediction for yield and selectivity can help chemists identify superior catalysts or reagents, providing essential AI support for rational reaction design. In polymer synthesis, AI application has also shown remarkable outcomes: (1) ML methods can establish quantitative relationship between polymer structures and properties, achieving accurate predictions and even target-oriented polymer design; (2) AI can aid and guide the design and optimization of polymerization processes, achieving end-to-end control and linking polymerization conditions directly to the products' functionalities; (3) AI application in biological macromolecules is equally thriving, predicting structures, designing functional sequences, and even autonomously performing closed-loop continuous directed evolution for proteins, RNA, and other macromolecules. Additionally, the advancement of automated experimentation paves the way for liberating synthetic chemists, significantly improving precision and efficiency in synthesis, work-up, isolation, and purification. This, coupled with AI's brainpower, heralds the advent of intelligent synthesis laboratories. These exciting developments demonstrate AI's substantial contribution to synthetic chemistry, signaling the dawn of an era of intelligent synthesis.

However, it is crucial to acknowledge that AI application in synthetic chemistry is still nascent, with challenges and limitations that cannot be overlooked. The quantity and quality of available open data in synthetic chemistry are far from satisfactory, lacking unbiased, large-scale datasets like ImageNet to support AI development. The digital representation of synthetic systems requires the improvements in standardization, interpretability, and applicability. Current molecular and reaction encodings lack standardized methods and deep chemical understanding, also posing challenges in encoding like the stochastic nature of polymer structures. The "black box" nature of existing AI models makes it difficult for chemists to comprehend the decision-making process, limiting the models' capacity to provide chemical insights. Issues with model availability also hinder the model application in new synthetic systems. To overcome these challenges and truly promote the healthy, sustainable development of AI synthetic chemistry, the following actions are recommended. First, enhancing data sharing and model openness. Following the FAIR (findable, accessible, interoperable, and reusable) principles, chemists should reshape the open data community of synthetic chemistry with advanced large models, so as to foster broader collaboration and innovation. Second, democratize AI, making the cutting-edge achievements of AI chemistry accessible to experimental chemists for frontline synthetic design. Third, focusing on software and hardware upgrades and optimization to better integrate AI technology and experimental synthetic processes. This will enable automated synthesis platforms to enter everyday laboratories and significantly enhance the efficiency and accuracy of synthetic experiments. It requires the joint efforts of chemists, computer scientists, and engineers to accelerate the process of intelligentization in synthetic chemistry.

With the continuous breakthroughs and development of ML and automation technologies, synthetic chemistry is undergoing a transformation from a traditional "manual" era to an "intelligent" era. In the near future, AI will play a vital role in every aspect of synthetic chemistry: (1) Molecular design. AI will use chemical databases and ML algorithms to design molecules with specific functions according to chemists' needs. (2) Synthetic pathway planning. For a given target molecule, AI models can efficiently plan synthetic pathways and provide detailed experimental schemes. (3) Experimental execution: Integrated AI with automated synthesis platforms/robots will create intelligent synthesis laboratories capable of conducting experiments, providing real-time feedback, and automatically adjusting experimental plans for condition optimization until the target product is synthesized with high selectivity and yield. (4) Remote interaction: AI systems deployed in the cloud enable chemists to interact remotely at any time and place *via* mobile phones/computers. (5) The commercialization of general

chemical databases, ML algorithms, and machine chemists will likely make AI and related automation technologies standard equipment in ordinary synthetic laboratories, greatly promoting the development of synthetic chemistry.

Acknowledgements This work was supported by the National Natural Science Foundation of China (22393890, You SL; 22393891 and 22031006, Luo S; 2203300, Pei J; 22371052, Chen M; 21991132, 21925102, 92056118, and 22331003, Zhang WB; 22331002 and 22125101, Lu H; 22071004, Mo F; 22393892 and 22071249, Liao K; 22122109 and 22271253, Hong X), the National Key R&D Program of China (2023YFF1205103, Pei J; 2020YFA0908100 and 2023YFF1204401, Zhang WB; 2022YFA1504301, Hong X), Zhejiang Provincial Natural Science Foundation of China (LDQ23B020002, Hong X), the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study (SN-ZJU-SIAS-006, Hong X), the CAS Youth Interdisciplinary Team (JCTD-2021-11, Hong X), Shenzhen Medical Research Fund (B2302037, Zhang WB), Beijing National Laboratory for Molecular Sciences (BNLMS-CXXM-202006, Zhang WB), the State Key Laboratory of Molecular Engineering of Polymers (Chen M), Haihe Laboratory of Sustainable Chemical Transformations and National Science & Technology Fundamental Resource Investigation Program of China (2023YFA1500008, Luo S).

Conflict of interest The authors declare no conflict of interest.

1 Zhang S, Roller S, Goyal N, Artetxe M, Chen M, Chen S, Dewan C, Diab M, Li X, Lin XV, Mihaylov T, Ott M, Shleifer S, Shuster K, Simig D, Koura PS, Sridhar A, Wang T, Zettlemoyer L. Opt: open pre-trained transformer language models. arXiv preprint, 2205.01068, 2022
2 Zhang Z, Gu Y, Han X, Chen S, Xiao C, Sun Z, Yao Y, Qi F, Guan J, Ke P, Cai Y, Zeng G, Tan Z, Liu Z, Huang M, Han W, Liu Y, Zhu X, Sun M. *AI Open*, 2021, 2: 216–224
3 Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. *J Mach Learn Res*, 2020, 21: 1–67
4 Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. Language models are few-shot learners. arXiv preprint, 2005.14165, 2020
5 https://chat.openai.com/
6 Guo T, Guo K, Nan B, Liang Z, Guo Z, Chawla NV, Wiest O, Zhang X. What can large language models do in chemistry? A comprehensive benchmark on eight tasks. arXiv preprint, 2305.18365, 2023
7 Boiko DA, MacKnight R, Kline B, Gomes G. *Nature*, 2023, 624: 570–578
8 Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, Hubert T, Baker L, Lai M, Bolton A, Chen Y, Lillicrap T, Hui F, Sifre L, van den Driessche G, Graepel T, Hassabis D. *Nature*, 2017, 550: 354–359
9 Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. *Nature*, 2021, 596: 583–589
10 Corey EJ, Long AK, Rubenstein SD. *Science*, 1985, 228: 408–418
11 Molga K, Szymkuć S, Grzybowski BA. *Acc Chem Res*, 2021, 54: 1094–1106
12 Mikulak-Klucznik B, Gołębiowska P, Bayly AA, Popik O, Klucznik T, Szymkuć S, Gajewska EP, Dittwald P, Staszewska-Krajewska O, Beker W, Badowski T, Scheidt KA, Molga K, Mlynarski J, Mrksich M, Grzybowski BA. *Nature*, 2020, 588: 83–88
13 Ucak UV, Ashyrmamatov I, Ko J, Lee J. *Nat Commun*, 2022, 13: 1186
14 Bragato M, von Rudorff GF, von Lilienfeld OA. *Chem Sci*, 2020, 11: 11859–11868
15 Ingold CK. *Chem Rev*, 1934, 15: 225–274
16 Kermack WO, Robinson R. *J Chem Soc Trans*, 1922, 121: 427–440
17 Hammett LP. *J Am Chem Soc*, 1937, 59: 96–103
18 Seeman JI. *Chem Rev*, 1983, 83: 83–134
19 Santiago CB, Guo JY, Sigman MS. *Chem Sci*, 2018, 9: 2398–2412
20 http://ibond.nankai.edu.cn/
21 Yang Q, Li Y, Yang JD, Liu Y, Zhang L, Luo S, Cheng JP. *Angew Chem Int Ed*, 2020, 59: 19282–19291
22 Peplow M. *Nature*, 2014, 512: 20–22
23 Giles J. *Nature*, 2012, 481: 430–431
24 Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcantara R, Darsow M, Guedj M, Ashburner M. *Nucleic Acids Res*, 2007, 36: D344–D350
25 Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP. *Nucleic Acids Res*, 2012, 40: D1100–D1107
26 Saito T, Kinugasa S. *Synth Engl Ed*, 2011, 4: 35–44
27 Gražulis S, Chateigner D, Downs RT, Yokochi AFT, Quirós M, Lutterotti L, Manakova E, Butkus J, Moeck P, Le Bail A. *J Appl Crystlogr*, 2009, 42: 726–729
28 John Wiley & Sons, Inc. Online spectral database: Quick access to millions of NMR, IR, Raman, UV-vis, and mass spectra. https://spectrabase.com/
29 Linstrom PJ, Mallard WG. *J Chem Eng Data*, 2001, 46: 1059–1063
30 Chambers J, Davies M, Gaulton A, Hersey A, Velankar S, Petryszak R, Hastings J, Bellis L, McGlinchey S, Overington JP. *J Cheminform*, 2013, 5: 3
31 Gallarati S, van Gerwen P, Laplaza R, Vela S, Fabrizio A, Corminboeuf C. *Chem Sci*, 2022, 13: 13782–13794
32 Irwin JJ, Tang KG, Young J, Dandarchuluun C, Wong BR, Khurelbaatar M, Moroz YS, Mayfield J, Sayle RA. *J Chem Inf Model*, 2020, 60: 6065–6073
33 Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH. *Nucleic Acids Res*, 2009, 37: W623–W633
34 Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, Wang J, Yu B, Zhang J, Bryant SH. *Nucleic Acids Res*, 2016, 44: D1202–D1213
35 Kearnes SM, Maser MR, Wleklinski M, Kast A, Doyle AG, Dreher SD, Hawkins JM, Jensen KF, Coley CW. *J Am Chem Soc*, 2021, 143: 18820–18826
36 Ramakrishnan R, Dral PO, Rupp M, von Lilienfeld OA. *Sci Data*, 2014, 1: 140022
37 Ruddigkeit L, van Deursen R, Blum LC, Reymond JL. *J Chem Inf Model*, 2012, 52: 2864–2875
38 Weininger D. *J Chem Inf Comput Sci*, 1988, 28: 31–36
39 Krenn M, Häse F, Nigam AK, Friederich P, Aspuru-Guzik A. *Mach Learn-Sci Technol*, 2020, 1: 045024
40 Schütt KT, Sauceda HE, Kindermans PJ, Tkatchenko A, Müller KR. *J Chem Phys*, 2018, 148: 241722
41 Lin TS, Coley CW, Mochigase H, Beech HK, Wang W, Wang Z, Woods E, Craig SL, Johnson JA, Kalow JA, Jensen KF, Olsen BD. *ACS Cent Sci*, 2019, 5: 1523–1531
42 Guo M, Shou W, Makatura L, Erps T, Foshey M, Matusik W. *Adv Sci*, 2022, 9: 2101864
43 Bender A, Schneider N, Segler M, Patrick Walters W, Engkvist O, Rodrigues T. *Nat Rev Chem*, 2022, 6: 428–442
44 Margraf JT. *Angew Chem Int Ed*, 2023, 62: e202219170
45 Lin Y, Zhang R, Wang D, Cernak T. *Science*, 2023, 379: 453–457
46 Rodríguez-Pérez R, Bajorath J. *J Med Chem*, 2020, 63: 8761–8777
47 Rodríguez-Pérez R, Bajorath J. *J Comput Aided Mol Des*, 2020, 34: 1013–1026

48 Ribeiro MT, Singh S, Guestrin C. Model-agnostic interpretability of machine learning. arXiv preprint, 1606.05386, 2019
49 Zhang WQ, Ge P, Jin WD, Guo J. Radar signal recognition based on TPOT and LIME. In: *2018 37th Chinese Control Conference (CCC)*. Wuhan, 2018. 4158–4163
50 Lu C, Liu Q, Wang C, Huang Z, Lin P, He L. Molecular property prediction: A multilevel quantum interactions modeling perspective. arXiv preprint, 1906.11081, 2019
51 Liu Y, Yang Q, Li Y, Zhang L, Luo S. *Chinese J Org Chem*, 2020, 40: 3812–3827
52 Fu Y, Liu L, Li RQ, Liu R, Guo QX. *J Am Chem Soc*, 2004, 126: 814–822
53 Alongi KS, Shields GC. Theoretical calculations of acid dissociation constants: a review article. In: Ralph W, Ed. *Annual Reports in Computational Chemistry*. Amsterdam: Elsevier, 2010. 113–138
54 Ho J, Coote ML. *WIREs Comput Mol Sci*, 2011, 1: 649–660
55 Seybold PG, Shields GC. *WIREs Comput Mol Sci*, 2015, 5: 290–297
56 Philipp DM, Watson MA, Yu HS, Steinbrecher TB, Bochevarov AD. *Int J Quantum Chem*, 2018, 118: e25561
57 Wu J, Kang Y, Pan P, Hou T. *Drug Discov Today*, 2022, 27: 103372
58 Jover J, Bosque R, Sales J. *QSAR Comb Sci*, 2007, 26: 385–397
59 Harding AP, Wedge DC, Popelier PLA. *J Chem Inf Model*, 2009, 49: 1914–1924
60 Jover J, Bosque R, Sales J. *QSAR Comb Sci*, 2008, 27: 1204–1215
61 Chen B, Zhang H, Li M. *Neural Comput Applic*, 2019, 31: 8297–8304
62 Zhou T, Jhamb S, Liang X, Sundmacher K, Gani R. *Chem Eng Sci*, 2018, 183: 95–105
63 Milletti F, Storchi L, Goracci L, Bendels S, Wagner B, Kansy M, Cruciani G. *Eur J Medicinal Chem*, 2010, 45: 4270–4279
64 Lu Y, Anand S, Shirley W, Gedeck P, Kelley BP, Skolnik S, Rodde S, Nguyen M, Lindvall M, Jia W. *J Chem Inf Model*, 2019, 59: 4706–4719
65 Fraczkiewicz R, Lobell M, Göller AH, Krenz U, Schoenneis R, Clark RD, Hillisch A. *J Chem Inf Model*, 2015, 55: 389–397
66 Roszak R, Beker W, Molga K, Grzybowski BA. *J Am Chem Soc*, 2019, 141: 17142–17149
67 Mayr F, Wieder M, Wieder O, Langer T. *Front Chem*, 2022, 10: 866585
68 Feng Y, Liu L, Wang JT, Huang H, Guo QX. *J Chem Inf Comput Sci*, 2003, 43: 2005–2013
69 Feng Y, Liu L, Wang JT, Zhao SW, Guo QX. *J Org Chem*, 2004, 69: 3129–3138
70 St. John PC, Guan Y, Kim Y, Etz BD, Kim S, Paton RS. *Sci Data*, 2020, 7: 244
71 Bosque R, Sales J. *J Chem Inf Comput Sci*, 2003, 43: 637–642
72 Xue CX, Zhang RS, Liu HX, Yao XJ, Liu MC, Hu ZD, Fan BT. *J Chem Inf Comput Sci*, 2004, 44: 669–677
73 Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V. *J Mol Graphics Model*, 2008, 27: 188–196
74 Xu Q, Xu J. *Monatsh Chem*, 2016, 148: 645–654
75 Nakajima M, Nemoto T. *Sci Rep*, 2021, 11: 20207
76 Wen M, Blau SM, Spotte-Smith EWC, Dwaraknath S, Persson KA. *Chem Sci*, 2021, 12: 1858–1868
77 Gao P, Zhang J, Qiu H, Zhao S. *Phys Chem Chem Phys*, 2021, 23: 13242–13249
78 Guo S, Jiang J, Ren H, Wang S. *J Phys Chem Lett*, 2023, 14: 7461–7468
79 Qu X, Latino DA, Aires-de-Sousa J. *J Cheminform*, 2013, 5: 34
80 Feng C, Sharman E, Ye S, Luo Y, Jiang J. *Sci China Chem*, 2019, 62: 1698–1703
81 St. John PC, Guan Y, Kim Y, Kim S, Paton RS. *Nat Commun*, 2020, 11: 2328
82 Li W, Luan Y, Zhang Q, Aires-de-Sousa J. *Mol Inf*, 2023, 42: e2200193
83 Unke OT, Meuwly M. *J Chem Theor Comput*, 2019, 15: 3678–3693
84 Shui Z, Karypis G. Heterogeneous molecular graph neural networks for predicting molecule properties. In: *Proceedings - 20th IEEE International Conference on Data Mining*. Sorrento, Italy: ICDM, 2020. 492–500
85 Simeon G, De Fabritiis G. TensorNet: cartesian tensor representations for efficient learning of molecular potentials. arXiv preprint, 2306.06482, 2023
86 Fang X, Liu L, Lei J, He D, Zhang S, Zhou J, Wang F, Wu H, Wang H. *Nat Mach Intell*, 2022, 4: 127–134
87 Zhou G, Gao Z, Ding Q, Zheng H, Xu H, Wei Z, Zhang L, Ke G. Uni-Mol: a universal 3D molecular representation learning framework. ChemRxiv preprint, 2023, doi: 10.26434/chemrxiv-2022-jjm0j-v4
88 Delaney JS. *J Chem Inf Comput Sci*, 2004, 44: 1000–1005
89 Mobley DL, Guthrie JP. *J Comput Aided Mol Des*, 2014, 28: 711–720
90 Hille C, Ringe S, Deimel M, Kunkel C, Acree WE, Reuter K, Oberhofer H. *J Chem Phys*, 2019, 150: 041710
91 Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, Leswing K, Pande V. *Chem Sci*, 2018, 9: 513–530
92 Tagade PM, Adiga SP, Park MS, Pandian S, Hariharan KS, Kolake SM. *J Phys Chem C*, 2018, 122: 11322–11333
93 Ghule S, Dash SR, Bagchi S, Joshi K, Vanka K. *ACS Omega*, 2022, 7: 11742–11755
94 Grambow CA, Pattanaik L, Green WH. *J Phys Chem Lett*, 2020, 11: 2992–2997
95 Jorner K, Brinck T, Norrby PO, Buttar D. *Chem Sci*, 2021, 12: 1163–1175
96 Houston PL, Nandi A, Bowman JM. *J Phys Chem Lett*, 2019, 10: 5250–5258
97 Liu Y, Yang Q, Cheng J, Zhang L, Luo S, Cheng JP. *ChemPhysChem*, 2023, 24: e202300162
98 Saini V, Sharma A, Nivatia D. *Phys Chem Chem Phys*, 2022, 24: 1821–1829
99 Orlandi M, Escudero-Casao M, Licini G. *J Org Chem*, 2021, 86: 3555–3564
100 Hoffmann G, Balcilar M, Tognetti V, Héroux P, Gaüzère B, Adam S, Joubert L. *J Comput Chem*, 2020, 41: 2124–2136
101 Cuesta SA, Moreno M, López RA, Mora JR, Paz JL, Márquez EA. *J Chem Inf Model*, 2023, 63: 507–521
102 Boobier S, Liu Y, Sharma K, Hose DRJ, Blacker AJ, Kapur N, Nguyen BN. *J Chem Inf Model*, 2021, 61: 4890–4899
103 Nie W, Liu D, Li S, Yu H, Fu Y. *J Chem Inf Model*, 2022, 62: 4319–4328
104 Käser S, Vazquez-Salazar LI, Meuwly M, Töpfer K. *Digital Discov*, 2023, 2: 28–58
105 Behler J, Parrinello M. *Phys Rev Lett*, 2007, 98: 146401
106 Han J, Zhang L, Car R, E W. *CiCP*, 2018, 23
107 Wang J, Olsson S, Wehmeyer C, Pérez A, Charron NE, de Fabritiis G, Noé F, Clementi C. *ACS Cent Sci*, 2019, 5: 755–767
108 Bowman JM, Qu C, Conte R, Nandi A, Houston PL, Yu Q. *J Chem Theor Comput*, 2023, 19: 1–17
109 Kang PL, Shang C, Liu ZP. *J Am Chem Soc*, 2019, 141: 20525–20536
110 Ouldridge TE. *Nat Comput*, 2018, 17: 3–29
111 van Speybroeck V, Gani R, Meier RJ. *Chem Soc Rev*, 2010, 39: 1764–1779
112 Bell RP. *The Proton in Chemistry*. Ithaca: Cornell University Press, 1973
113 Stewart R. *The Proton: Applications to Organic Chemistry*. Orlando: Academic Press, 1985
114 Xue XS, Ji P, Zhou B, Cheng JP. *Chem Rev*, 2017, 117: 8622–8648
115 Yang JD, Ji P, Xue XS, Cheng JP. *J Am Chem Soc*, 2018, 140: 8611–8623
116 Yang JD, Xue J, Cheng JP. *Chem Soc Rev*, 2019, 48: 2913–2926
117 Cai Z, Liu T, Lin Q, He J, Lei X, Luo F, Huang Y. *J Chem Inf Model*, 2023, 63: 2936–2947
118 Wei W, Hogues H, Sulea T. *J Chem Inf Model*, 2023, 63: 5169–5181

119 Luo YR. *Comprehensive Handbook of Chemical Bond Energies*. Boca Raton: CRC Press, 2007
120 Rupp M, Tkatchenko A, Müller KR, von Lilienfeld OA. *Phys Rev Lett*, 2012, 108: 058301
121 Ramakrishnan R, Hartmann M, Tapavicza E, von Lilienfeld OA. *J Chem Phys*, 2015, 143: 084111
122 Blum LC, Reymond JL. *J Am Chem Soc*, 2009, 131: 8732–8733
123 Di Martino RMC, Maxwell BD, Pirali T. *Nat Rev Drug Discov*, 2023, 22: 562–584
124 Meanwell NA. *J Agric Food Chem*, 2023, 71: 18087–18122
125 Bhat V, Welin ER, Guo X, Stoltz BM. *Chem Rev*, 2017, 117: 4528–4561
126 Mahboob I, Shafiq I, Shafique S, Akhter P, Amjad US, Hussain M, Park YK. *Chem Eng J*, 2022, 441: 136063
127 Jorner K, Tomberg A, Bauer C, Sköld C, Norrby PO. *Nat Rev Chem*, 2021, 5: 240–255
128 Hughes ED. *Nature*, 1942, 149: 126–130
129 Komp E, Janulaitis N, Valleau S. *Phys Chem Chem Phys*, 2022, 24: 2692–2705
130 Greaves TL, Schaffarczyk McHale KS, Burkart-Radke RF, Harper JB, Le TC. *Phys Chem Chem Phys*, 2021, 23: 2742–2752
131 Mayr H, Patz M. *Angew Chem Int Ed Engl*, 1994, 33: 938–957
132 Mayr H, Ofial AR. *SAR QSAR Environ Res*, 2015, 26: 619–646
133 Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AWR, Bridgland A, Penedones H, Petersen S, Simonyan K, Crossan S, Kohli P, Jones DT, Silver D, Kavukcuoglu K, Hassabis D. *Nature*, 2020, 577: 706–710
134 Westermayr J, Marquetand P. *Mach Learn-Sci Technol*, 2020, 1: 043001
135 Unke OT, Chmiela S, Sauceda HE, Gastegger M, Poltavsky I, Schütt KT, Tkatchenko A, Müller KR. *Chem Rev*, 2021, 121: 10142–10186
136 Noé F, Tkatchenko A, Müller KR, Clementi C. *Annu Rev Phys Chem*, 2020, 71: 361–390
137 Schaaf LL, Fako E, De S, Schäfer A, Csányi G. *npj Comput Mater*, 2023, 9: 180
138 Chmiela S, Vassilev-Galindo V, Unke OT, Kabylda A, Sauceda HE, Tkatchenko A, Müller KR. *Sci Adv*, 2023, 9: eadf0873
139 Shu Y, Varga Z, Jasper A, Espinosa-Garcia J, Corchado JC, Truhlar DG. *Comput Phys Commun*, 2024, 294: 108937
140 Schlegel HB. *J Comput Chem*, 2003, 24: 1514–1527
141 Wei GF, Liu ZP. *J Chem Theor Comput*, 2016, 12: 4698–4706
142 Huang SD, Shang C, Zhang XJ, Liu ZP. *Chem Sci*, 2017, 8: 6327–6337
143 Zhang XJ, Shang C, Liu ZP. *J Chem Phys*, 2017, 147: 152706
144 Zhang XJ, Shang C, Liu ZP. *Phys Chem Chem Phys*, 2017, 19: 4725–4733
145 Fang YH, Ma SC, Liu ZP. *J Phys Chem C*, 2019, 123: 19347–19353
146 Corey EJ, Wipke WT. *Science*, 1969, 166: 178–192
147 Segler MHS, Waller MP. *Chem Eur J*, 2017, 23: 5966–5971
148 Segler MHS, Preuss M, Waller MP. *Nature*, 2018, 555: 604–610
149 Kishimoto A, Buesser B, Chen B, Botea A. Depth-first proof-number search with heuristic edge cost and application to chemical synthesis planning. In: Advances in Neural Information Processing Systems 32 (Nips 2019). Vancouver, 2019. 7226–7236
150 Chen B, Li C, Dai H, Song L. Learning retrosynthetic planning with neural guided A* Search. In: *Proceedings of the 37th International Conference on Machine Learning*. online, 2020. 119: 1608–1616
151 Xie S, Yan R, Han P, Xia Y, Wu L, Guo C, Yang B, Qin T. Retrographer: retrosynthetic planning with graph search. arXiv preprint, 2206.11477, 2022
152 Kim J, Ahn S, Lee H, Shin J. Self-improved retrosynthetic planning. arXiv preprint, 2106.04880, 2021
153 Yu Y, Wei Y, Kuang K, Huang Z, Yao H, Wu F. GRASP: navigating retrosynthetic planning with goal-driven policy. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, 2022. 10257–10268
154 Liu G, Xue D, Xie S, Xia Y, Tripp A, Maziarz K, Segler M, Qin T, Zhang Z, Liu TY. Retrosynthetic planning with dual value networks. arXiv preprint, 2301.13755, 2023
155 Coley CW, Thomas Iii DA, Lummiss JAM, Jaworski JN, Breen CP, Schultz V, Hart T, Fishman JS, Rogers L, Gao H, Hicklin RW, Plehiers PP, Byington J, Piotti JS, Green WH, Hart AJ, Jamison TF, Jensen KF. *Science*, 2019, 365: eaax1566
156 Lin Y, Zhang Z, Mahjour B, Wang D, Zhang R, Shim E, McGrath A, Shen Y, Brugger N, Turnbull R, Trice S, Jasty S, Cernak T. *Nat Commun*, 2021, 12: 7327
157 Wołos A, Koszelewski D, Roszak R, Szymkuć S, Moskal M, Ostaszewski R, Herrera BT, Maier JM, Brezicki G, Samuel J, Lummiss JAM, McQuade DT, Rogers L, Grzybowski BA. *Nature*, 2022, 604: 668–676
158 Jewett JC, Bertozzi CR. *Chem Soc Rev*, 2010, 39: 1272–1279
159 Park Y, Kim Y, Chang S. *Chem Rev*, 2017, 117: 9247–9301
160 Braconi E. *Nat Rev Methods Primers*, 2023, 3: 74
161 Rinehart NI, Zahrt AF, Henle JJ, Denmark SE. *Acc Chem Res*, 2021, 54: 2041–2054
162 Żurański AM, Martinez Alvarado JI, Shields BJ, Doyle AG. *Acc Chem Res*, 2021, 54: 1856–1865
163 Muratov EN, Bajorath J, Sheridan RP, Tetko IV, Filimonov D, Poroikov V, Oprea TI, Baskin II, Varnek A, Roitberg A, Isayev O, Curtalolo S, Fourches D, Cohen Y, Aspuru-Guzik A, Winkler DA, Agrafiotis D, Cherkasov A, Tropsha A. *Chem Soc Rev*, 2020, 49: 3525–3564
164 Zahrt AF, Athavale SV, Denmark SE. *Chem Rev*, 2020, 120: 1620–1689
165 Oliveira JCA, Frey J, Zhang SQ, Xu LC, Li X, Li SW, Hong X, Ackermann L. *Trends Chem*, 2022, 4: 863–885
166 Crawford JM, Kingston C, Toste FD, Sigman MS. *Acc Chem Res*, 2021, 54: 3136–3148
167 Strieth-Kalthoff F, Sandfort F, Segler MHS, Glorius F. *Chem Soc Rev*, 2020, 49: 6154–6168
168 Yang L, Zhu L, Zhang S, Hong X. *Chin J Chem*, 2022, 40: 2106–2117
169 Sandfort F, Strieth-Kalthoff F, Kühnemund M, Beecks C, Glorius F. *Chem*, 2020, 6: 1379–1390
170 Maley SM, Kwon DH, Rollins N, Stanley JC, Sydora OL, Bischof SM, Ess DH. *Chem Sci*, 2020, 11: 9665–9674
171 Gallarati S, Fabregat R, Laplaza R, Bhattacharjee S, Wodrich MD, Corminboeuf C. *Chem Sci*, 2021, 12: 6879–6889
172 Moskal M, Beker W, Szymkuć S, Grzybowski BA. *Angew Chem Int Ed*, 2021, 60: 15230–15235
173 Li B, Su S, Zhu C, Lin J, Hu X, Su L, Yu Z, Liao K, Chen H. *J Cheminform*, 2023, 15: 72
174 Tsuji N, Sidorov P, Zhu C, Nagata Y, Gimadiev T, Varnek A, List B. *Angew Chem Int Ed*, 2023, 62: e202218659
175 Xu Y, Gao Y, Su L, Wu H, Tian H, Zeng M, Xu C, Zhu X, Liao K. *Angew Chem Int Ed*, 2023, 62
176 Shields BJ, Stevens J, Li J, Parasram M, Damani F, Alvarado JIM, Janey JM, Adams RP, Doyle AG. *Nature*, 2021, 590: 89–96
177 Gensch T, Smith SR, Colacot TJ, Timsina YN, Xu G, Glasspoole BW, Sigman MS. *ACS Catal*, 2022, 12: 7773–7780
178 Liles JP, Rouget-Virbel C, Wahlman JLH, Rahimoff R, Crawford JM, Medlin A, O'Connor VS, Li J, Roytman VA, Toste FD, Sigman MS. *Chem*, 2023, 9: 1518–1537
179 van Dijk L, Haas BC, Lim NK, Clagg K, Dotson JJ, Treacy SM, Piechowicz KA, Roytman VA, Zhang H, Toste FD, Miller SJ, Gosselin F, Sigman MS. *J Am Chem Soc*, 2023, 145: 20959–20967
180 Ahneman DT, Estrada JG, Lin S, Dreher SD, Doyle AG. *Science*, 2018, 360: 186–190
181 Qiu J, Xie J, Su S, Gao Y, Meng H, Yang Y, Liao K. *Chem*, 2022, 8: 3275–3287
182 Saebi M, Nan B, Herr JE, Wahlers J, Guo Z, Zurański AM, Kogej T, Norrby PO, Doyle AG, Chawla NV, Wiest O. *Chem Sci*, 2023, 14: 4997–5005
183 Schwaller P, Vaucher AC, Laino T, Reymond JL. *Mach Learn-Sci*

Technol, 2021, 2: 015016
184 Strieth-Kalthoff F, Sandfort F, Kühnemund M, Schäfer FR, Kuchen H, Glorius F. *Angew Chem Int Ed*, 2022, 61: e202204647
185 Gallegos LC, Luchini G, St. John PC, Kim S, Paton RS. *Acc Chem Res*, 2021, 54: 827–836
186 Singh S, Sunoj RB. *Acc Chem Res*, 2023, 56: 402–412
187 Zahrt AF, Henle JJ, Rose BT, Wang Y, Darrow WT, Denmark SE. *Science*, 2019, 363: eaau5631
188 Reid JP, Sigman MS. *Nature*, 2019, 571: 343–348
189 Zhang X, Chung LW, Wu YD. *Acc Chem Res*, 2016, 49: 1302–1310
190 Zhang S, Xu L, Li S, Oliveira JCA, Li X, Ackermann L, Hong X. *Chem Eur J*, 2023, 29: e202202834
191 Xu L, Zhang S, Li X, Tang M, Xie P, Hong X. *Angew Chem Int Ed*, 2021, 60: 22804–22811
192 Xu LC, Frey J, Hou X, Zhang SQ, Li YY, Oliveira JCA, Li SW, Ackermann L, Hong X. *Nat Synth*, 2023, 2: 321–330
193 Zhang ZJ, Li SW, Oliveira JCA, Li Y, Chen X, Zhang SQ, Xu LC, Rogge T, Hong X, Ackermann L. *Nat Commun*, 2023, 14: 3149
194 Beker W, Gajewska EP, Badowski T, Grzybowski BA. *Angew Chem Int Ed*, 2019, 58: 4515–4519
195 Melville J, Hargis C, Davenport MT, Hamilton RS, Ess DH. *J Phys Org Chem*, 2022, 35: e4405
196 Pesciullesi G, Schwaller P, Laino T, Reymond JL. *Nat Commun*, 2020, 11: 4874
197 Li X, Zhang S, Xu L, Hong X. *Angew Chem Int Ed*, 2020, 59: 13253–13259
198 Guan Y, Coley CW, Wu H, Ranasinghe D, Heid E, Struble TJ, Pattanaik L, Green WH, Jensen KF. *Chem Sci*, 2021, 12: 2198–2208
199 Boni YT, Cammarota RC, Liao K, Sigman MS, Davies HML. *J Am Chem Soc*, 2022, 144: 15549–15561
200 Dhawa U, Tian C, Wdowik T, Oliveira JCA, Hao J, Ackermann L. *Angew Chem Int Ed*, 2020, 59: 13451–13457
201 Caldeweyher E, Elkin M, Gheibi G, Johansson M, Sköld C, Norrby PO, Hartwig JF. *J Am Chem Soc*, 2023, 145: 17367–17376
202 Beker W, Roszak R, Wołos A, Angello NH, Rathore V, Burke MD, Grzybowski BA. *J Am Chem Soc*, 2022, 144: 4819–4827
203 Häse F, Roch LM, Kreisbeck C, Aspuru-Guzik A. *ACS Cent Sci*, 2018, 4: 1134–1145
204 Singh S, Pareek M, Changotra A, Banerjee S, Bhaskararao B, Balamurugan P, Sunoj RB. *Proc Natl Acad Sci USA*, 2020, 117: 1339–1345
205 Guo Y, He X, Su Y, Dai Y, Xie M, Yang S, Chen J, Wang K, Zhou D, Wang C. *J Am Chem Soc*, 2021, 143: 5755–5762
206 Raccuglia P, Elbert KC, Adler PDF, Falk C, Wenny MB, Mollo A, Zeller M, Friedler SA, Schrier J, Norquist AJ. *Nature*, 2016, 533: 73–76
207 Bai Y, Wilbraham L, Slater BJ, Zwijnenburg MA, Sprick RS, Cooper AI. *J Am Chem Soc*, 2019, 141: 9063–9071
208 Doan Tran H, Kim C, Chen L, Chandrasekaran A, Batra R, Venkatram S, Kamal D, Lightstone JP, Gurnani R, Shetty P, Ramprasad M, Laws J, Shelton M, Ramprasad R. *J Appl Phys*, 2020, 128: 171104
209 Kim C, Chandrasekaran A, Huan TD, Das D, Ramprasad R. *J Phys Chem C*, 2018, 122: 17575–17585
210 Pilania G, Wang C, Jiang X, Rajasekaran S, Ramprasad R. *Sci Rep*, 2013, 3: 2810
211 Wu K, Sukumar N, Lanzillo NA, Wang C, "Rampi" Ramprasad R, Ma R, Baldwin AF, Sotzing G, Breneman C. *J Polym Sci Part B-Polym Phys*, 2016, 54: 2082–2091
212 Simine L, Allen TC, Rossky PJ. *Proc Natl Acad Sci USA*, 2020, 117: 13945–13948
213 Wang Y, Xie T, France-Lanord A, Berkley A, Johnson JA, Shao-Horn Y, Grossman JC. *Chem Mater*, 2020, 32: 4144–4151
214 Aldeghi M, Coley CW. *Chem Sci*, 2022, 13: 10486–10498
215 Jha A, Chandrasekaran A, Kim C, Ramprasad R. *Model Simul Mater Sci Eng*, 2019, 27: 024002
216 Sun W, Zheng Y, Yang K, Zhang Q, Shah AA, Wu Z, Sun Y, Feng L, Chen D, Xiao Z, Lu S, Li Y, Sun K. *Sci Adv*, 2019, 5: eaay4275
217 Xu C, Wang Y, Barati Farimani A. *npj Comput Mater*, 2023, 9: 64
218 Kuenneth C, Ramprasad R. *Nat Commun*, 2023, 14: 4099
219 Rahman A, Deshpande P, Radue MS, Odegard GM, Gowtham S, Ghosh S, Spear AD. *Compos Sci Tech*, 2021, 207: 108627
220 Park J, Shim Y, Lee F, Rammohan A, Goyal S, Shim M, Jeong C, Kim DS. *ACS Polym Au*, 2022, 2: 213–222
221 Yamada H, Liu C, Wu S, Koyama Y, Ju S, Shiomi J, Morikawa J, Yoshida R. *ACS Cent Sci*, 2019, 5: 1717–1730
222 Venkatram S, Batra R, Chen L, Kim C, Shelton M, Ramprasad R. *J Phys Chem B*, 2020, 124: 6046–6054
223 Kuenneth C, Rajan AC, Tran H, Chen L, Kim C, Ramprasad R. *Patterns*, 2021, 2: 100238
224 Mannodi-Kanakkithodi A, Pilania G, Huan TD, Lookman T, Ramprasad R. *Sci Rep*, 2016, 6: 20952
225 Afzal MAF, Haghighatlari M, Ganesh SP, Cheng C, Hachmann J. *J Phys Chem C*, 2019, 123: 14610–14618
226 Wu S, Kondo Y, Kakimoto M, Yang B, Yamada H, Kuwajima I, Lambard G, Hongo K, Xu Y, Shiomi J, Schick C, Morikawa J, Yoshida R. *npj Comput Mater*, 2019, 5: 66
227 Kim C, Batra R, Chen L, Tran H, Ramprasad R. *Comput Mater Sci*, 2021, 186: 110067
228 Zhou T, Wu Z, Chilukoti HK, Müller-Plathe F. *J Chem Theor Comput*, 2021, 17: 3772–3782
229 Martin TB, Audus DJ. *ACS Polym Au*, 2023, 3: 239–258
230 Rubens M, Vrijsen JH, Laun J, Junkers T. *Angew Chem Int Ed*, 2019, 58: 3183–3187
231 Rubens M, Van Herck J, Junkers T. *ACS Macro Lett*, 2019, 8: 1437–1441
232 Zhang B, Mathoor A, Junkers T. *Angew Chem Int Ed*, 2023, 62: e202308838
233 Knox ST, Parkinson SJ, Wilding CYP, Bourne RA, Warren NJ. *Polym Chem*, 2022, 13: 1576–1585
234 Rizkin BA, Shkolnik AS, Ferraro NJ, Hartman RL. *Nat Mach Intell*, 2020, 2: 200–209
235 Gu Y, Lin P, Zhou C, Chen M. *Sci China Chem*, 2021, 64: 1039–1046
236 Zhao B, Cheng J, Gao J, Haddleton DM, Wilson P. *Macro Chem Phys*, 2023, 224: 2300039
237 Chapman R, Gormley AJ, Herpoldt KL, Stevens MM. *Macromolecules*, 2014, 47: 8541–8547
238 Chapman R, Gormley AJ, Stenzel MH, Stevens MM. *Angew Chem Int Ed*, 2016, 128: 4500–4503
239 Liu Z, Lv Y, An Z. *Angew Chem Int Ed*, 2017, 56: 13852–13856
240 Li R, Zhang S, Li Q, Qiao GG, An Z. *Angew Chem Int Ed*, 2022, 61: e202213396
241 Enciso AE, Fu L, Russell AJ, Matyjaszewski K. *Angew Chem Int Ed*, 2018, 57: 933–936
242 Xu J, Jung K, Atme A, Shanmugam S, Boyer C. *J Am Chem Soc*, 2014, 136: 5508–5519
243 Gormley AJ, Yeow J, Ng G, Conway Ó, Boyer C, Chapman R. *Angew Chem Int Ed*, 2018, 57: 1557–1562
244 Zhou Y, Gu C, Zheng L, Shan F, Chen G. *Polym Chem*, 2022, 13: 989–996
245 Zheng Y, Luo Y, Feng K, Zhang W, Chen G. *ACS Macro Lett*, 2019, 8: 326–330
246 Jafari VF, Mossayebi Z, Allison-Logan S, Shabani S, Qiao GG. *Chem Eur J*, 2023, 29: e202301767
247 Lv C, He C, Pan X. *Angew Chem Int Ed*, 2018, 57: 9430–9433
248 Stubbs C, Congdon T, Davis J, Lester D, Richards SJ, Gibson MI. *Macromolecules*, 2019, 52: 7603–7612
249 Zhong Y, Zeberl BJ, Wang X, Luo J. *Acta Biomater*, 2018, 73: 21–37
250 Ladmiral V, Mantovani G, Clarkson GJ, Cauet S, Irwin JL, Haddleton DM. *J Am Chem Soc*, 2006, 128: 4823–4830
251 Yan Y, Liu L, Xiong H, Miller JB, Zhou K, Kos P, Huffman KE, Elkassih S, Norman JW, Carstens R, Kim J, Minna JD, Siegwart DJ. *Proc Natl Acad Sci USA*, 2016, 113: E5702–E5710
252 Upadhya R, Kanagala MJ, Gormley AJ. *Macromol Rapid Commun*,

2019, 40: 1900528
253 Hughes I, Hunter D. *Curr Opin Chem Biol*, 2001, 5: 243–247
254 Schmatloch S, Meier MAR, Schubert US. *Macromol Rapid Commun*, 2003, 24: 33–46
255 Zhang S, Wang L, Fu X. *Sci Sin Chim*, 2023, 53: 3–8
256 Bannigan P, Bao Z, Hickman RJ, Aldeghi M, Häse F, Aspuru-Guzik A, Allen C. *Nat Commun*, 2023, 14: 35
257 Kumar R, Le N, Tan Z, Brown ME, Jiang S, Reineke TM. *ACS Nano*, 2020, 14: 17626–17639
258 Ye S, Meftahi N, Lyskov I, Tian T, Whitfield R, Kumar S, Christofferson AJ, Winkler DA, Shih CJ, Russo S, Leroux JC, Bao Y. *Chem*, 2023, 9: 924–947
259 Fransen KA, Av-Ron SHM, Buchanan TR, Walsh DJ, Rota DT, Van Note L, Olsen BD. *Proc Natl Acad Sci USA*, 2023, 120: e2220021120
260 Sanders MA, Chittari SS, Sherman N, Foley JR, Knight AS. *J Am Chem Soc*, 2023, 145: 9686–9692
261 Gormley AJ, Webb MA. *Nat Rev Mater*, 2021, 6: 642–644
262 Pollice R, dos Passos Gomes G, Aldeghi M, Hickman RJ, Krenn M, Lavigne C, Lindner-D'Addario M, Nigam AK, Ser CT, Yao Z, Aspuru-Guzik A. *Acc Chem Res*, 2021, 54: 849–860
263 Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, Li B, Madabhushi A, Shah P, Spitzer M, Zhao S. *Nat Rev Drug Discov*, 2019, 18: 463–477
264 Reis M, Gusev F, Taylor NG, Chung SH, Verber MD, Lee YZ, Isayev O, Leibfarth FA. *J Am Chem Soc*, 2021, 143: 17677–17689
265 Kosuri S, Borca CH, Mugnier H, Tamasi M, Patel RA, Perez I, Kumar S, Finkel Z, Schloss R, Cai L, Yarmush ML, Webb MA, Gormley AJ. *Adv Healthcare Mater*, 2022, 11: 2102101
266 Tamasi MJ, Patel RA, Borca CH, Kosuri S, Mugnier H, Upadhya R, Murthy NS, Webb MA, Gormley AJ. *Adv Mater*, 2022, 34: 2201809
267 Wu G, Zhou H, Zhang J, Tian ZY, Liu X, Wang S, Coley CW, Lu H. *Nat Synth*, 2023, 2: 515–526
268 Bordin N, Dallago C, Heinzinger M, Kim S, Littmann M, Rauer C, Steinegger M, Rost B, Orengo C. *Trends Biochem Sci*, 2023, 48: 345–359
269 Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, Millán C, Park H, Adams C, Glassman CR, DeGiovanni A, Pereira JH, Rodrigues AV, van Dijk AA, Ebrecht AC, Opperman DJ, Sagmeister T, Buhlheller C, Pavkov-Keller T, Rathinaswamy MK, Dalwadi U, Yip CK, Burke JE, Garcia KC, Grishin NV, Adams PD, Read RJ, Baker D. *Science*, 2021, 373: 871–876
270 Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, dos Santos Costa A, Fazel-Zarandi M, Sercu T, Candido S, Rives A. *Science*, 2023, 379: 1123–1130
271 Yu T, Cui H, Li JC, Luo Y, Jiang G, Zhao H. *Science*, 2023, 379: 1358–1363
272 van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, Söding J, Steinegger M. *Nat Biotechnol*, 2024, 42: 243–246
273 Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM. *Nat Methods*, 2019, 16: 1315–1322
274 Madani A, Krause B, Greene ER, Subramanian S, Mohr BP, Holton JM, Olmos Jr. JL, Xiong C, Sun ZZ, Socher R, Fraser JS, Naik N. *Nat Biotechnol*, 2023, 41: 1099–1106
275 Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J, Fergus R. *Proc Natl Acad Sci USA*, 2021, 118: e2016239118
276 Jehl P, Manguy J, Shields DC, Higgins DG, Davey NE. *Nucleic Acids Res*, 2016, 44: W11–W15
277 Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, Gibbs T, Feher T, Angerer C, Steinegger M, Bhowmik D, Rost B. *IEEE Trans Pattern Anal Mach Intell*, 2022, 44: 7112–7127
278 Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. *Bioinformatics*, 2022, 38: 2102–2110
279 Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, Kaufman KW, Renfrew PD, Smith CA, Sheffler W, Davis IW, Cooper S, Treuille A, Mandell DJ, Richter F, Ban Y-EA, Fleishman SJ, Corn JE, Kim DE, Lyskov S, Berrondo M, Mentzer S, Popović Z, Havranek JJ, Karanicolas J, Das R, Meiler J, Kortemme T, Gray JJ, Kuhlman B, Baker D, Bradley P. *Meth Enzymol*, 2011, 487: 545–574
280 Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. *Nucleic Acids Res*, 2005, 33: W382–W388
281 Liu Y, Zhang L, Wang W, Zhu M, Wang C, Li F, Zhang J, Li H, Chen Q, Liu H. *Nat Comput Sci*, 2022, 2: 451–462
282 Dauparas J, Anishchenko I, Bennett N, Bai H, Ragotte RJ, Milles LF, Wicky BIM, Courbet A, de Haas RJ, Bethel N, Leung PJY, Huddy TF, Pellock S, Tischer D, Chan F, Koepnick B, Nguyen H, Kang A, Sankaran B, Bera AK, King NP, Baker D. *Science*, 2022, 378: 49–56
283 Huang B, Xu Y, Hu X, Liu Y, Liao S, Zhang J, Huang C, Hong J, Chen Q, Liu H. *Nature*, 2022, 602: 523–528
284 Anishchenko I, Pellock SJ, Chidyausiku TM, Ramelot TA, Ovchinnikov S, Hao J, Bafna K, Norn C, Kang A, Bera AK, DiMaio F, Carter L, Chow CM, Montelione GT, Baker D. *Nature*, 2021, 600: 547–552
285 Watson JL, Juergens D, Bennett NR, Trippe BL, Yim J, Eisenach HE, Ahern W, Borst AJ, Ragotte RJ, Milles LF, Wicky BIM, Hanikel N, Pellock SJ, Courbet A, Sheffler W, Wang J, Venkatesh P, Sappington I, Torres SV, Lauko A, De Bortoli V, Mathieu E, Ovchinnikov S, Barzilay R, Jaakkola TS, DiMaio F, Baek M, Baker D. *Nature*, 2023, 620: 1089–1100
286 Ingraham JB, Baranov M, Costello Z, Barber KW, Wang W, Ismail A, Frappier V, Lord DM, Ng-Thow-Hing C, Van Vlack ER, Tie S, Xue V, Cowles SC, Leung A, Rodrigues JV, Morales-Perez CL, Ayoub AM, Green R, Puentes K, Oplinger F, Panwar NV, Obermeyer F, Root AR, Beam AL, Poelwijk FJ, Grigoryan G. *Nature*, 2023, 623: 1070–1078
287 Jiang L, Althoff EA, Clemente FR, Doyle L, Rothlisberger D, Zanghellini A, Gallaher JL, Betker JL, Tanaka F, Barbas Iii CF, Hilvert D, Houk KN, Stoddard BL, Baker D. *Science*, 2008, 319: 1387–1391
288 Yeh AHW, Norn C, Kipnis Y, Tischer D, Pellock SJ, Evans D, Ma P, Lee GR, Zhang JZ, Anishchenko I, Coventry B, Cao L, Dauparas J, Halabiya S, DeWitt M, Carter L, Houk KN, Baker D. *Nature*, 2023, 614: 774–780
289 Yang KK, Wu Z, Arnold FH. *Nat Methods*, 2019, 16: 687–694
290 Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CPI, Springer M, Sander C, Marks DS. *Nat Biotechnol*, 2017, 35: 128–135
291 Shamsi Z, Chan M, Shukla D. *J Phys Chem B*, 2020, 124: 3845–3854
292 Hsu C, Nisonoff H, Fannjiang C, Listgarten J. *Nat Biotechnol*, 2022, 40: 1114–1122
293 Biswas S, Khimulya G, Alley EC, Esvelt KM, Church GM. *Nat Methods*, 2021, 18: 389–396
294 Riesselman AJ, Ingraham JB, Marks DS. *Nat Methods*, 2018, 15: 816–822
295 Esvelt KM, Carlson JC, Liu DR. *Nature*, 2011, 472: 499–503
296 Dickinson BC, Leconte AM, Allen B, Esvelt KM, Liu DR. *Proc Natl Acad Sci USA*, 2013, 110: 9007–9012
297 Thuronyi BW, Koblan LW, Levy JM, Yeh WH, Zheng C, Newby GA, Wilson C, Bhaumik M, Shubina-Oleinik O, Holt JR, Liu DR. *Nat Biotechnol*, 2019, 37: 1070–1079
298 Yu T, Boob AG, Singh N, Su Y, Zhao H. *Cell Syst*, 2023, 14: 633–644
299 Enghiad B, Xue P, Singh N, Boob AG, Shi C, Petrov VA, Liu R, Peri SS, Lane ST, Gaither ED, Zhao H. *Nat Commun*, 2022, 13: 2697
300 HamediRad M, Chao R, Weisberg S, Lian J, Sinha S, Zhao H. *Nat Commun*, 2019, 10: 5150
301 Sato K, Akiyama M, Sakakibara Y. *Nat Commun*, 2021, 12: 941
302 Taubert O, von der Lehr F, Bazarova A, Faber C, Knechtges P, Weiel M, Debus C, Coquelin D, Basermann A, Streit A, Kesselheim S, Götz M, Schug A. *Commun Biol*, 2023, 6: 913
303 Wang W, Feng C, Han R, Wang Z, Ye L, Du Z, Wei H, Zhang F, Peng Z, Yang J. *Nat Commun*, 2023, 14: 7266
304 Yao W, Xiong DC, Yang Y, Geng C, Cong Z, Li F, Li BH, Qin X, Wang LN, Xue WY, Yu N, Zhang H, Wu X, Liu M, Ye XS. *Nat*

Synth, 2022, 1: 854–863
305 Fang G, Lin D, Liao K. *Chin J Chem*, 2023, 41: 1075–1079
306 Alvarado-Urbina G, Sathe GM, Liu WC, Gillen MF, Duck PD, Bender R, Ogilvie KK. *Science*, 1981, 214: 270–274
307 Caruthers MH. *Science*, 1985, 230: 281–285
308 Merrifield RB, Stewart JM. *Nature*, 1965, 207: 522–523
309 Merrifield RB. *Science*, 1965, 150: 178–185
310 Scaringe SA, Wincott FE, Caruthers MH. *J Am Chem Soc*, 1998, 120: 11820–11821
311 Plante OJ, Palmacci ER, Seeberger PH. *Science*, 2001, 291: 1523–1527
312 Legrand M, Foucard A. *J Chem Educ*, 1978, 55: 767
313 Deadman BJ, Battilocchio C, Sliwinski E, Ley SV. *Green Chem*, 2013, 15: 2050–2055
314 Sutherland JD, Tu NP, Nemcek TA, Searle PA, Hochlowski JE, Djuric SW, Pan JY. *SLAS Tech*, 2014, 19: 176–182
315 Tu NP, Searle PA, Sarris K. *SLAS Tech*, 2016, 21: 459–469
316 Durrer J, Agrawal P, Ozgul A, Neuhauss SCF, Nama N, Ahmed D. *Nat Commun*, 2022, 13: 6370
317 Plutschack MB, Pieber B, Gilmore K, Seeberger PH. *Chem Rev*, 2017, 117: 11796–11893
318 Guidi M, Seeberger PH, Gilmore K. *Chem Soc Rev*, 2020, 49: 8910–8932
319 Machida K, Hirose Y, Fuse S, Sugawara T, Takahashi T. *Chem Pharm Bull*, 2010, 58: 87–93
320 Mijalis AJ, Thomas Iii DA, Simon MD, Adamo A, Beaumont R, Jensen KF, Pentelute BL. *Nat Chem Biol*, 2017, 13: 464–466
321 Perera D, Tucker JW, Brahmbhatt S, Helal CJ, Chong A, Farrell W, Richardson P, Sach NW. *Science*, 2018, 359: 429–434
322 Li C, Callahan AJ, Simon MD, Totaro KA, Mijalis AJ, Phadke KS, Zhang G, Hartrampf N, Schissel CK, Zhou M, Zong H, Hanson GJ, Loas A, Pohl NLB, Verhoeven DE, Pentelute BL. *Nat Commun*, 2021, 12: 4396
323 Li C, Callahan AJ, Phadke KS, Bellaire B, Farquhar CE, Zhang G, Schissel CK, Mijalis AJ, Hartrampf N, Loas A, Verhoeven DE, Pentelute BL. *ACS Cent Sci*, 2021, 8: 205–213
324 Li C, Zhang G, Mohapatra S, Callahan AJ, Loas A, Gómez-Bombarelli R, Pentelute BL. *Adv Sci*, 2022, 9: 2201988
325 Steiner S, Wolf J, Glatzel S, Andreou A, Granda JM, Keenan G, Hinkley T, Aragon-Camarasa G, Kitson PJ, Angelone D, Cronin L. *Science*, 2019, 363: eaav2211
326 Osipyan A, Shaabani S, Warmerdam R, Shishkina SV, Boltz H, Dömling A. *Angew Chem Int Ed*, 2020, 59: 12423–12427
327 Sagmeister P, Lebl R, Castillo I, Rehrl J, Kruisz J, Sipek M, Horn M, Sacher S, Cantillo D, Williams JD, Kappe CO. *Angew Chem Int Ed*, 2021, 60: 8139–8148
328 Ahn GN, Sharma BM, Lahore S, Yim SJ, Vidyacharan S, Kim DP. *Commun Chem*, 2021, 4: 53
329 Chatterjee S, Guidi M, Seeberger PH, Gilmore K. *Nature*, 2020, 579: 379–384
330 Nandiwale KY, Hart T, Zahrt AF, Nambiar AMK, Mahesh PT, Mo Y, Nieves-Remacha MJ, Johnson MD, García-Losada P, Mateos C, Rincón JA, Jensen KF. *React Chem Eng*, 2022, 7: 1315–1327
331 Amara Z, Bellamy JFB, Horvath R, Miller SJ, Beeby A, Burgard A, Rossen K, Poliakoff M, George MW. *Nat Chem*, 2015, 7: 489–495
332 Bana P, Örkényi R, Lövei K, Lakó Á, Túrós GI, Éles J, Faigl F, Greiner I. *BioOrg Medicinal Chem*, 2017, 25: 6180–6189
333 Bana P, Szigetvári Á, Kóti J, Éles J, Greiner I. *React Chem Eng*, 2019, 4: 652–657
334 Baranczak A, Tu NP, Marjanovic J, Searle PA, Vasudevan A, Djuric SW. *ACS Med Chem Lett*, 2017, 8: 461–465
335 Thomson CG, Banks C, Allen M, Barker G, Coxon CR, Lee AL, Vilela F. *J Org Chem*, 2021, 86: 14079–14094
336 Li J, Ballmer SG, Gillis EP, Fujii S, Schmidt MJ, Palazzolo AME, Lehmann JW, Morehouse GF, Burke MD. *Science*, 2015, 347: 1221–1226
337 Li J, Grillo AS, Burke MD. *Acc Chem Res*, 2015, 48: 2297–2307
338 Blair DJ, Chitti S, Trobe M, Kostyra DM, Haley HMS, Hansen RL, Ballmer SG, Woods TJ, Wang W, Mubayi V, Schmidt MJ, Pipal RW, Morehouse GF, Palazzolo Ray AME, Gray DL, Gill AL, Burke MD. *Nature*, 2022, 604: 92–97
339 Lehmann JW, Blair DJ, Burke MD. *Nat Rev Chem*, 2018, 2: 0115
340 Hwang YJ, Coley CW, Abolhasani M, Marzinzik AL, Koch G, Spanka C, Lehmann H, Jensen KF. *Chem Commun*, 2017, 53: 6649–6652
341 Baumgartner LM, Dennis JM, White NA, Buchwald SL, Jensen KF. *Org Process Res Dev*, 2019, 23: 1594–1601
342 Mo Y, Rughoobur G, Nambiar AMK, Zhang K, Jensen KF. *Angew Chem Int Ed*, 2020, 59: 20890–20894
343 Mo Y, Lu Z, Rughoobur G, Patil P, Gershenfeld N, Akinwande AI, Buchwald SL, Jensen KF. *Science*, 2020, 368: 1352–1357
344 Sun AC, Steyer DJ, Allen AR, Payne EM, Kennedy RT, Stephenson CRJ. *Nat Commun*, 2020, 11: 6202
345 van Putten R, Eyke NS, Baumgartner LM, Schultz VL, Filonenko GA, Jensen KF, Pidko EA. *ChemSusChem*, 2022, 15: e202200333
346 Debrouwer W, Kimpe W, Dangreau R, Huvaere K, Gemoets HPL, Mottaghi M, Kuhn S, van Aken K. *Org Process Res Dev*, 2020, 24: 2319–2325
347 Adamo A, Beingessner RL, Behnam M, Chen J, Jamison TF, Jensen KF, Monbaliu JCM, Myerson AS, Revalor EM, Snead DR, Stelzer T, Weeranoppanant N, Wong SY, Zhang P. *Science*, 2016, 352: 61–67
348 Jiang T, Bordi S, McMillan AE, Chen KY, Saito F, Nichols PL, Wanner BM, Bode JW. *Chem Sci*, 2021, 12: 6977–6982
349 McMillan AE, Wu WWX, Nichols PL, Wanner BM, Bode JW. *Chem Sci*, 2022, 13: 14292–14299
350 Kitson PJ, Marie G, Francoia JP, Zalesskiy SS, Sigerson RC, Mathieson JS, Cronin L. *Science*, 2018, 359: 314–319
351 Hou W, Bubliauskas A, Kitson PJ, Francoia JP, Powell-Davies H, Gutierrez JMP, Frei P, Manzano JS, Cronin L. *ACS Cent Sci*, 2021, 7: 212–218
352 Bubliauskas A, Blair DJ, Powell-Davies H, Kitson PJ, Burke MD, Cronin L. *Angew Chem Int Ed*, 2022, 61: e202116108
353 Manzano JS, Hou W, Zalesskiy SS, Frei P, Wang H, Kitson PJ, Cronin L. *Nat Chem*, 2022, 14: 1311–1318
354 Skilton RA, Bourne RA, Amara Z, Horvath R, Jin J, Scully MJ, Streng E, Tang SLY, Summers PA, Wang J, Pérez E, Asfaw N, Aydos GLP, Dupont J, Comak G, George MW, Poliakoff M. *Nat Chem*, 2015, 7: 1–5
355 Fitzpatrick DE, Battilocchio C, Ley SV. *Org Process Res Dev*, 2016, 20: 386–394
356 Roch LM, Häse F, Kreisbeck C, Tamayo-Mendoza T, Yunker LPE, Hein JE, Aspuru-Guzik A. *Sci Robot*, 2018, 3: eaat5559
357 Rohrbach S, Šiaučiulis M, Chisholm G, Pirvan PA, Saleeb M, Mehr SHM, Trushina E, Leonov AI, Keenan G, Khan A, Hammer A, Cronin L. *Science*, 2022, 377: 172–180
358 Li J, Tu Y, Liu R, Lu Y, Zhu X. *Adv Sci*, 2020, 7: 1901957
359 Burger B, Maffettone PM, Gusev VV, Aitchison CM, Bai Y, Wang X, Li X, Alston BM, Li B, Clowes R, Rankin N, Harris B, Sprick RS, Cooper AI. *Nature*, 2020, 583: 237–241
360 Zhu Q, Zhang F, Huang Y, Xiao H, Zhao LY, Zhang XC, Song T, Tang XS, Li X, He G, Chong BC, Zhou JY, Zhang YH, Zhang B, Cao JQ, Luo M, Wang S, Ye GL, Zhang WJ, Chen X, Cong S, Zhou D, Li H, Li J, Zou G, Shang WW, Jiang J, Luo Y. *Natl Sci Rev*, 2022, 9: nwac190
361 Ley SV, Fitzpatrick DE, Ingham RJ, Myers RM. *Angew Chem Int Ed*, 2015, 54: 3449–3464
362 O'Brien M, Koos P, Browne DL, Ley SV. *Org Biomol Chem*, 2012, 10: 7031
363 O'Brien AG, Horváth Z, Lévesque F, Lee JW, Seidel-Morgenstern A, Seeberger PH. *Angew Chem Int Ed*, 2012, 51: 7028–7030
364 Ingham RJ, Battilocchio C, Fitzpatrick DE, Sliwinski E, Hawkins JM, Ley SV. *Angew Chem Int Ed*, 2015, 54: 144–148
365 Granda JM, Donina L, Dragone V, Long DL, Cronin L. *Nature*, 2018, 559: 377–381

366 Zeng L, Burton L, Yung K, Shushan B, Kassel DB. *J Chromatogr A*, 1998, 794: 3–13
367 Koppitz M, Brailsford A, Wenz M. *J Comb Chem*, 2005, 7: 714–720
368 Guth O, Krewer D, Freudenberg B, Paulitz C, Hauser M, Ilg K. *J Comb Chem*, 2008, 10: 875–882
369 Xu H, Lin J, Liu Q, Chen Y, Zhang J, Yang Y, Young MC, Xu Y, Zhang D, Mo F. *Chem*, 2022, 8: 3202–3214
370 Xu H, Lin J, Zhang D, Mo F. *Nat Commun*, 2023, 14: 3095
371 Herges R. *J Chem Inf Comput Sci*, 1990, 30: 377–383
372 Caramelli D, Granda JM, Mehr SHM, Cambié D, Henson AB, Cronin L. *ACS Cent Sci*, 2021, 7: 1821–1830
373 McNally A, Prier CK, MacMillan DWC. *Science*, 2011, 334: 1114–1117
374 Troshin K, Hartwig JF. *Science*, 2017, 357: 175–181
375 Zahrt AF, Mo Y, Nandiwale KY, Shprints R, Heid E, Jensen KF. *J Am Chem Soc*, 2022, 144: 22599–22610
376 Lan T, An Q. *J Am Chem Soc*, 2021, 143: 16804–16812
377 Yoon J, Cao Z, Raju RK, Wang Y, Burnley R, Gellman AJ, Barati Farimani A, Ulissi ZW. *Mach Learn-Sci Technol*, 2021, 2: 045018
378 Aldoseri A, Al-Khalifa KN, Hamouda AM. *Appl Sci*, 2023, 13: 7082
379 Baum ZJ, Yu X, Ayala PY, Zhao Y, Watkins SP, Zhou Q. *J Chem Inf Model*, 2021, 61: 3197–3212
380 Tetko IV, Engkvist O, Koch U, Reymond J-, Chen H. *Mol Inf*, 2016, 35: 615–621
381 Emami FS, Vahid A, Wylie EK, Szymkuć S, Dittwald P, Molga K, Grzybowski BA. *Angew Chem Int Ed*, 2015, 54: 10797–10801
382 Angello NH, Rathore V, Beker W, Wołos A, Jira ER, Roszak R, Wu TC, Schroeder CM, Aspuru-Guzik A, Grzybowski BA, Burke MD. *Science*, 2022, 378: 399–405
383 Jia X, Lynch A, Huang Y, Danielson M, Lang′at I, Milder A, Ruby AE, Wang H, Friedler SA, Norquist AJ, Schrier J. *Nature*, 2019, 573: 251–255
384 Huerta EA, Blaiszik B, Brinson LC, Bouchard KE, Diaz D, Doglioni C, Duarte JM, Emani M, Foster I, Fox G, Harris P, Heinrich L, Jha S, Katz DS, Kindratenko V, Kirkpatrick CR, Lassila-Perini K, Madduri RK, Neubauer MS, Psomopoulos FE, Roy A, Rübel O, Zhao Z, Zhu R. *Sci Data*, 2023, 10: 487
385 Yano J, Gaffney KJ, Gregoire J, Hung L, Ourmazd A, Schrier J, Sethian JA, Toma FM. *Nat Rev Chem*, 2022, 6: 357–370
386 Liu J, Hein JE. *Nat Synth*, 2023, 2: 464–466
387 Chithrananda S, Grand G, Ramsundar B. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. arXiv preprint, 2010.09885, 2020
388 Ross J, Belgodere B, Chenthamarakshan V, Padhi I, Mroueh Y, Das P. *Nat Mach Intell*, 2022, 4: 1256–1264
389 Frey NC, Soklaski R, Axelrod S, Samsi S, Gómez-Bombarelli R, Coley CW, Gadepally V. *Nat Mach Intell*, 2023, 5: 1297–1305
390 Wang X, Jiang S, Hu W, Ye S, Wang T, Wu F, Yang L, Li X, Zhang G, Chen X, Jiang J, Luo Y. *J Am Chem Soc*, 2022, 144: 16069–16076
391 Bédard AC, Adamo A, Aroh KC, Russell MG, Bedermann AA, Torosian J, Yue B, Jensen KF, Jamison TF. *Science*, 2018, 361: 1220–1225