RSC Advances

REVIEW

View Article Online
View Journal | View Issue

Cite this: *RSC Adv.*, 2026, **16**, 27008

Deep learning mechanism and its application in biomacromolecules

Hairui Li,[a] Guocheng Zhang,[*b] Bin Yan,[b] Zaihang Ye,[cde] Xudong Zhen[f] and Yang Liu[*g]

Biomacromolecules are pivotal to the advancement of designing functional systems for biomedical and biomaterial applications. Artificial intelligence (AI), especially deep neural networks, is revolutionizing these domains, driving the transition from predictive to generative design. This review surveys the mechanisms and performance of these cutting-edge structure-predicting deep-learning models and their generative counterparts used for the *de novo* design of biomacromolecules. AI tools can also be used to predict self-assembly behavior, design high-affinity binders, and pave the way for novel structures with customized functions. This review concludes with a discussion on the current advantages, long-term barriers, and exciting directions for future exploration. AI is shifting discovery from a slow, trial-and-error process to a rapid, strategic, data-driven paradigm, redefining the search for new biological and material concepts as a targeted endeavor.

Received 21st January 2026
Accepted 6th May 2026

DOI: 10.1039/d6ra00550k

rsc.li/rsc-advances

1. Introduction

Biomacromolecules are high-molecular-weight organic polymers, such as proteins, nucleic acids, and polysaccharides, which function as the fundamental, structural and functional agents of biological systems. These molecules exhibit complex hierarchical architectures, determined by their monomeric sequences and stabilized by non-covalent interactions, including π–π interactions, hydrogen bonding, and hydrophobic effects. These interactions drive them into thermodynamically stable conformations, designed for specific biological activities. For example, peptides are short chains consisting of two to thousands of amino acids, and they adopt primary to quaternary structures. They can fold into proteins and serve as versatile molecular tools, orchestrating a vast array of processes in living organisms.[1] Their ability to function as natural receptor ligands, facilitate the transmembrane transport of molecules, and exhibit antimicrobial properties has promoted their application in fields including pharmacology, therapeutics, and immunology.[2–4] In the broad aspect of biomedical engineering and materials science, biomacromolecules are characterized not by their biological origin but by their capacity to artificially synthesize their structure and their capacity to form "functional systems". They serve as supramolecular assemblies that execute programmable tasks. For example, synthesized peptides are thought of as versatile "programmable LEGO bricks", which can be structurally modified by engineered amino acid side chains, designed sequences and stereoisomeric compositions. These factors define the hierarchical architectures of biomacromolecules, establish their function, and act in various roles from enzymatic catalysis to the structural support of the cytoskeleton and molecular recognition in immune response.[5–7]

Machine learning, particularly deep neural networks, has achieved significant prominence in recent years. Breakthroughs in novel architectures, such as the attention mechanism and diffusion models, are profoundly influencing new domains. Their application in biological and biomaterial research is accelerating experimental workflows and boosting discovery rates by thousands or even billions of times.[8–10] In this review, we present a comprehensive survey of the architectures, advantages and applications of attention-based models (*e.g.*, Transformers) and generative models (*e.g.* diffusion models and variational autoencoders (VAEs)). Moreover, we critically discuss the use of such state-of-the-art methods for predicting and *de novo* designing proteins and peptide-based materials.

2. Definition and characteristics of transformer architecture

We refer to the large language model (LLM) as a family of deep neural network models primarily based on a Transformer

[a]*Department of Plastic and cosmetic surgery, West China Tianfu Hospital, Sichuan University, Chengdu 610213, China*
[b]*School of Mechanical and Aviation Manufacturing Engineering, Anyang Institute of Technology, Anyang 455000, China. E-mail: zgc_uestc@163.com*
[c]*CAS Key Laboratory for Biological Effects of Nanomaterials & Nanosafety, Laboratory of Theoretical and Computational Nanoscience, National Center for Nanoscience and Technology, Beijing 100190, China*
[d]*Sino-Danish Center for Education and Research, Beijing 101408, China*
[e]*University of Chinese Academy of Sciences, Beijing 100049, China*
[f]*The First People's Hospital of Shihezi City, Shihezi 832061, China*
[g]*College of Polymer Science and Engineering, State Key Laboratory of Polymer Materials Engineering, Sichuan University, Chengdu 610065, China*

architecture[11]—typically comprising tens of billions of parameters and trained on massive text corpora. The specific architecture is illustrated in Fig. 1(a). The most distinguishing feature of the transformer architecture is the self-attention mechanism that fills a key design limitation of prior sequential models, such as recurrent neural networks (RNNs).[15] In conventional RNNs, the input sequence is fed to the RNNs sequentially, during which the historical context information is all reduced to one fixed-size hidden state vector. Thus, RNNs suffer from a bottleneck where later input context is less likely to be captured, and long-term dependencies are compromised. The attention mechanism can be seen as a workaround of such a bottleneck by allowing the model to explicitly select and weigh all the elements in the input sentence space equally. Instead of being conditioned on a condensed representation, the model can dynamically determine and consider the most relevant pieces of input at each step of the output generation process, permitting an improved awareness of contextual relations between words regardless of their distance.[16–19] Moreover, the self-attention operation is parallelizable—unlike RNNs, which process sequentially—enabling highly parallelized computation. This design is well-suited for modern hardware accelerators (*e.g.*, GPUs and TPUs), facilitating significantly faster training on large datasets and enabling the training of models at a scale commensurate with today's LLMs.

A wealth of empirical research, including the scaling laws of Kaplan *et al.* (referred to as KM) and the Chinchilla framework,[20,21] confirms that LLMs exhibit superior capabilities, a trend governed by predictable scaling laws. These laws allow one to extrapolate the performance of large models, measured for instance in latency, from results obtained for relatively small models. Additionally, they assist in tackling two bottlenecks in LLM research: first, the full-scale testing of new models is often impractical. Scaling laws address this by allowing researchers to extrapolate performance from small, affordable experiments, enabling a more efficient allocation of computing resources. Second, the long, expensive training phases of LLMs are susceptible to instabilities, *e.g.*, training loss spikes. Here, scaling laws play a diagnostic role: they enable practitioners to monitor training health by comparing current performance to predicted scaling curves, helping flag unexpected behaviors or potential failures.

Another key characteristic of LLMs is their emergent abilities—formally defined by Fedus *et al.* as skills that arise unpredictably and exclusively in models of a certain scale.[22] These abilities are characterized by a discontinuous jump in performance—a sharp transition from random chance to above-random competence—upon crossing a threshold in model size or computational budget. Three notable examples of these emergent capabilities are as follows. (1) In-context learning (ICL): first demonstrated with GPT-3, ICL enables models to tackle new tasks by implicitly inferring patterns from prompts containing natural language instructions and examples—without any gradient-based updates.[23] This emergent capability scales with model size; for instance, it is robustly present in 175 B-parameter GPT-3 but absent in its relatively small predecessors. It also remains highly sensitive to both model size and task characteristics.[22] (2) Instruction following: instruction following is learned by instruction tuning (*i.e.* fine-tuning on a mix of various tasks that are framed as natural language instructions). LLMs can use this to generalize well to the held-out tasks described through instructions. Performance is found to be heavily scale sensitive: for example, a model of size 68 B showed dramatic improvements in performance on held-out tasks; for a model with 8 B or fewer parameters, there was rarely any improvement observed.[24,25] (3) Chain-of-thought (CoT): LLMs

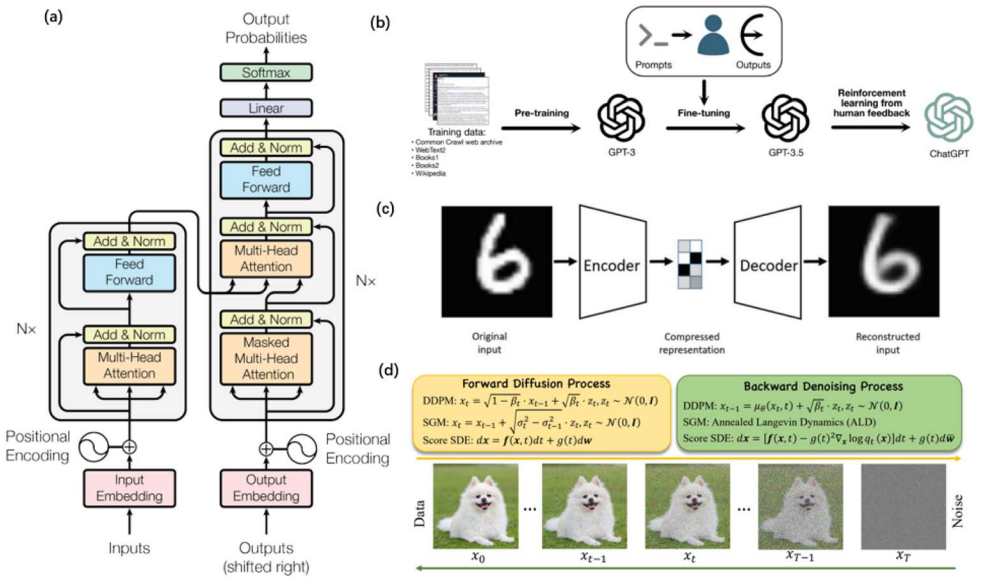

Fig. 1 Examples of the (a) transformer architecture[11] (copied with permission), (b) finetuning LLM[12] (copied with permission) (copyright 2025, Springer Nature), (c) VAE process[13] (copied with permission) and (d) diffusion model[14] (diffusion and denoising process of DDPM, SGM and Score SDE) (copied with permission, copyright 2025, ACM, Inc.).

can reason across multiple steps (*e.g.*, math problems) using CoT, which requires a model to explicitly generate a series of reasoning steps. CoT is particularly effective in models with more than 60–100 B parameters; we theorize that pre-training data with code might explain this. The magnitude of gains is wide-ranging across tasks and benchmarks.[26,27]

3. Training process of LLM and its expansion to other areas

Training an LLM typically involves the following three steps: pretraining, fine-tuning and reinforcement learning, as shown in Fig. 1(b). Pretraining is the building phase that enables the language model to obtain some basic language knowledge and text generation skills by training models on a large corpus.[23,28] The amount and quality of the pretraining data are key factors in deciding the ultimate capabilities of the model. In addition, successful pretraining demands well-designed model architecture, efficient training acceleration schemes, and advanced optimization algorithms.

Beyond pretraining, LLMs are often fine-tuned to enhance performance and adaptability, typically through two stages: instruction tuning (or supervised fine-tuning) and alignment tuning. The former aims to either bring out or reinforce some of the model's potential, for example, the ability to follow instructions, by training the model with exemplars of tasks. In addition to instruction tuning, alignment tuning involves training the model so that its behaviors in task and non-task scenarios match human values and preferences, for example, by inducing helpfulness, truthfulness, or harmlessness. We primarily implement alignment tuning *via* high-quality, deliberately curated training data to induct ethical and normative guidelines in how a model responds.

Reinforcement learning from human feedback (RLHF): this is a popular fine-tuning approach used for the value alignment of LLMs by optimizing human-defined criteria, such as helpfulness, honesty and harmlessness, with the help of reinforcement learning algorithms and human-provided feedback.[29,30] We note that the typical RLHF pipeline has two primary elements: a trained reward model (RM) based on a set of human preferences and a reinforcement learning algorithm (*e.g.* proximal policy optimization (PPO)[31]). PPO finetunes the language model with rewards, relying on the output of the RM. RM, which produces scalar estimates of human preferences over generated text, is trained *via* fine-tuning a previously trained language model or learning from scratch on a human-annotated training set of comparisons. The RM is built by having the seed LM generate multiple responses to a prompt and ordering these responses *via* human annotations to create a calibration set. This ordering removes annotator noise, and the RM is trained with supervised learning to model the reward-preferring outcome. For instance, InstructGPT employed this method with a 6 B parameter GPT-3 model as the reward model, training it to approximate human ranking.[32]

Even though the attention algorithm was originally designed for LLMs, it can also be tuned and leveraged to predict various biomacromolecule and biomaterial properties (*e.g.*, self-assembly dynamics or the design of high-affinity binders) by extracting complex patterns from training data.[33–35] The underlying process is analogous to next-token prediction in LLMs. We will illustrate this with concrete examples in subsequent sections.

4. Architecture of a generative model

A generative model is an AI algorithm trained to produce novel, synthetic data that reflect the underlying patterns and distribution of its training dataset. Variational autoencoder (VAE) is a family of generative models that use a hybrid approach of neural network and graphical model to estimate an unknown or complex data distribution in an unsupervised way.[36,37] Its architecture is shown in Fig. 1(c). VAE is a robust framework for unsupervised learning, which provides data with probabilistic interpretation and creates controlled data from the latent space. VAE is based on the Helmholtz Machine, where the true data distribution is learned by modeling simple distributions in the latent space as projections onto the complex data manifold. The primary objective is to reconstruct the distribution over the training data to generate new data samples through sampling. VAEs benefit from using standard function approximation and stochastic gradient descent for training, which facilitate their ease of training. This ease-of-training has allowed VAEs to be applied in a wide range of applications, such as generative models, semi-supervised learning and representation learning. Their usage can also be extended to downstream tasks, including image analysis, image generation, motion prediction, and zero-shot learning scenarios.

Diffusion models are probabilistic generative models that learn how to deconvolute the degradation process applied to the training data using a two-phase pipeline: forward diffusion, in which more and more noise is added to the training data (*e.g.*, gradually convolving images with Gaussian filters), and reverse denoising, in which the denoising is iterated towards the original data, as illustrated in Fig. 1(d). During inference, these models iteratively refine random noise through incremental transformations, with the neural network predicting denoising adjustments at each step. We categorize them under three fundamental formulations (Fig. 1(d)): denoising diffusion probabilistic models (DDPMs),[38,39] score-based generative models (SGMs),[40,41] and stochastic differential equations (Score SDEs),[42,43] which present complementary understanding of the generative process and together dominate the current investigation in this area.

5. Machine learning application in structure and interaction prediction

A decades-long effort has been dedicated to precisely predict the structure of biomacromolecules (*e.g.*, proteins) from the monomer (*e.g.*, amino acid) sequence, which has been a major challenge in molecular biology. The biological activities of these biomacromolecules are determined by their exact spatial

configuration. Solving this problem using deep learning would contribute to unraveling the disease process, as well as drug design and the elimination of drug-resistant pathogens. One of the primary challenges lies in ensuring the quality of the training data, as this fundamentally determines the performance of deep learning architectures. Accordingly, we survey the most famous databases in this domain. These include BioGRID,[46–48] which catalogues genetic and chemical protein interactions, HPRD,[49–52] a centralized hub for human protein-disease associations, and UniProt,[53–56] which provides annotated protein sequences. Furthermore, we consider DIP,[57–59] known for its experimentally curated protein–protein interactions, and STRING,[60,61] which aggregates both known and predicted protein–protein interactions with functional associations. These databases have served as the foundation for the development of numerous machine learning models.[62–71]

However, we primarily focus on state-of-the-art deep learning models, particularly those built upon the attention algorithm. Using the same datasets but has the potential to explore more complex relationships between input and output variables, especially within large-scale data exhibiting long-range dependencies, AlphaFold2 has revolutionized our understanding of biological systems by illuminating the molecular foundations of life with unprecedented clarity.[44] It leverages a two-staged network architecture that combines linear sequence information with two-dimensional distance maps, as shown in Fig. 2(a). By refining atomic positions with accuracy through an SE(3)-equivariant Transformer architecture, it predicts structures with quality comparable to those of high-end X-ray crystallography and cryo-electron microscopy techniques. Similarly, RoseTTAFold employs an attention module within its multi-layer architecture to integrate information over one-dimensional sequence, two-dimensional distance maps and three-dimensional spatial coordinates,[45] as illustrated in Fig. 2(b). It reaches the same accuracy as DeepMind CASP14, solving X-ray and cryo-EM structure challenges and also explaining unknown protein functions in minutes. It is found that this network can also directly predict protein–protein complex structures from a protein sequence, bypassing the need for subunit modeling or docking. Beyond protein structures, RoseTTAFold All-Atom (RFAA) integrates the residue-level representations of amino acids and DNA bases with the atomic-level descriptions of small molecules, metal ions, and chemical modifications. This unified approach enables the modeling of multi-component complexes—including proteins, nucleic acids, and ligands—directly from sequence and structural data. Chatterjee *et al.* proposed PepMLM, a peptide design tool grounded on the protein language model, ESM2, which employs a masked language modeling (MLM) strategy.[72] In MLM, a model learns deep semantic characteristics by predicting randomly masked tokens within a given sequence. It was trained on 10 000 known peptide–protein interaction pairs, with binding peptides systematically masked. It was later evaluated on a test set of 203 peptide–protein interaction pairs, thus showing that it is also capable of producing new binding peptides. Therefore, protein structure prediction and biomolecular interactions have long been central research goals. However, machine learning now faces challenges far beyond basic tasks, such as single-point mutation analysis. The focus has shifted to modeling vastly more complex systems, from engineered peptides and noncanonical side chains to peptide–synthetic polymer interactions. Nevertheless, the scarcity of high-quality databases of noncanonical engineered biomacromolecules—due to the immense chemical space and limited high-resolution structures—hinders the development of robust models for accurate classification.

Although AlphaFold2 and the RoseTTAFold series have developed a new paradigm in protein structure prediction, they still face some inherent limitations in capturing the dynamic and heterogeneous characteristics of protein systems. AlphaFold2 and RoseTTAFold are primarily optimized for predicting the static, folded structures of natural proteins, often encountering challenges in deciphering intrinsically disordered regions, the structural consequences of point mutations, and the conformational ensembles crucial for elucidating protein misfolding. While RFAA significantly expands the prediction scope to encompass non-protein ligands, metal ions, and certain post-translational modifications, it remains constrained in predicting the subtle physical properties of non-natural amino acids or highly flexible coacervate systems. Additionally, these machine learning models generally cannot account for environmental factors, such as pH, ionic strength, and solvent effects, which also play important roles in protein structure and interaction predictions. These constraints stem from the scarcity of high-quality databases for noncanonical engineered biomacromolecules; given the vast chemical space and the paucity of high-resolution structures, the development of robust models for accurate classification is hindered.

The application of LLMs in biopolymers (including non-standard poly amino acids) also constitutes a frontier research field, evolving from traditional manual feature extraction to an end-to-end "chemical language" model that applies Transformer architectures for exploring massive chemical spaces. To effectively utilize LLMs, researchers represent structural notations, such as simplified molecular-input line-entry system (SMILES) strings, as a specialized language, allowing LLMs to learn grammatical and syntactical rules through unsupervised pretraining on massive datasets—such as the 100 million virtual strings for polyBERT[73] and about one million structure–property combinations for polyTAO.[74] Strategic preprocessing plays an important role in this process. The workflow entails canonicalizing the polymer-simplified molecular-input line-entry system to ensure structural uniqueness, employing character-level tokenization to lower vocabulary size while accurately capturing complex heterocyclic motifs, and implementing self-supervised tasks, such as token masking, to derive robust latent representations, commonly referred to as "fingerprints". Through multitask deep neural networks, these learned fingerprints can be connected to many physical, thermal, and mechanical properties. This approach delivers accuracy close to those of conventional methods while operating at speeds up to two orders of magnitude higher. Additionally, advanced frameworks, such as PolyNC,[75] combine LLM prompts with "chemical language" to construct language-to-

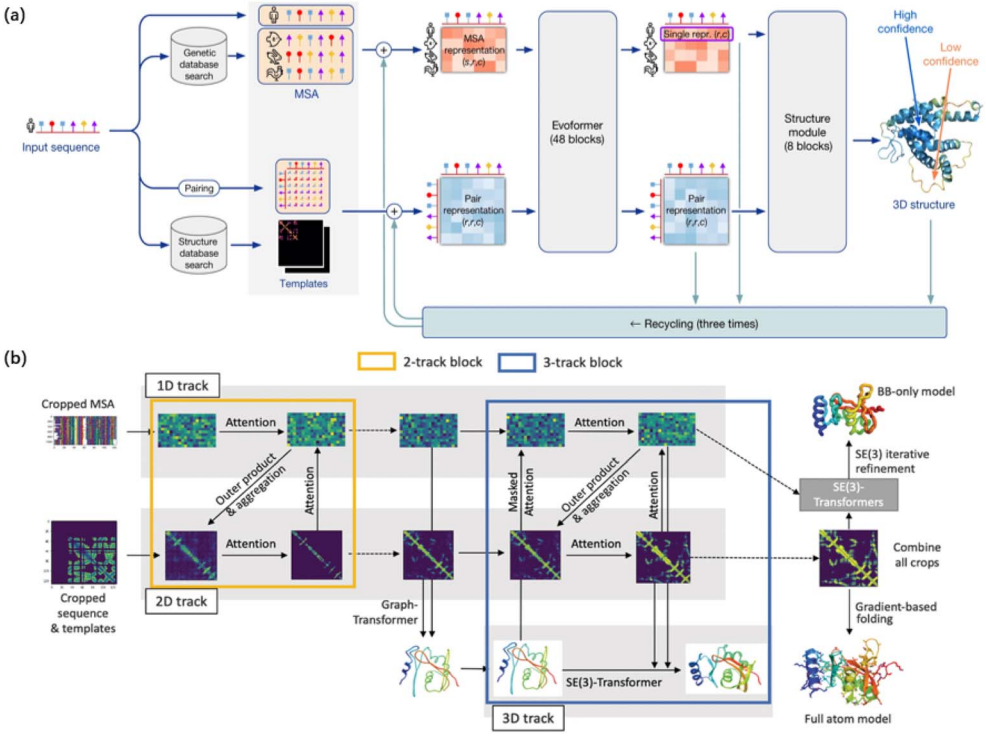

Fig. 2 Architectures of (a) AlphaFold model combining linear sequence information with two-dimensional distance maps (copied with permission,[44] copyright 2021, John Jumper *et al.*) and (b) RoseTTAFold model with 1D, 2D, and 3D attention tracks (copied with permission,[45] copyright 2025, the American Association for the Advancement of Science).

property agents that support unified multitask learning. Within a single model, these agents can perform both regression for quantitative property prediction and classification based on material attributes, for example, categorizing materials by their heat resistance class. Furthermore, LLMs allow on-demand reverse design by inverting the pipeline: property-to-SMILES translation allows models to generate novel, chemically valid structures, with success rates exceeding 99%, which satisfy specific user-defined target values. Existing literature highlights that developed architectures, such as transformers and LLMs, perform excellently in exploring large chemical spaces, yielding high validity (64.7%) and uniqueness (89.6%) in applications, such as vanadyl catalyst optimization.[76,77] Nevertheless, their architectural complexity often exhibits limited interpretability. To deal with this problem, explainable AI methods, including SHapley Additive exPlanations (SHAP), MolAnchor, and counterfactual explanations, should be utilized to extract interpretable feature patterns and map them to chemically meaningful understandings. Therefore, for biomacromolecular modeling, a two-step strategy should be used: first, pre-training LLMs on large-scale chemical databases to learn sequence–structure relationships through chemical language modeling; second, combining control calculations and explainable AI-based structural analysis to avoid invalid statistical correlations and ensure correct and reasonable biological motif learning. In this way, combining attention mechanisms-based LLMs can enable scalable, interpretable, and high-speed prediction for the discovery and rational design of complex biomacromolecules.

Beyond peptide sequences, machine learning models are increasingly used to investigate the relationships between the structure and properties of nucleic acids, including both chemically synthesized oligonucleotides and biologically derived DNA sequences. They are applied to predict important biophysical behaviors, thereby allowing the accurate design of functional genetic elements and therapeutic aptamers. Using high-density DNA microarray gene expression data, Won *et al.* leveraged machine learning algorithms, including a multilayer perceptron and *k*-nearest neighbor, together with a majority-voting ensemble classifier, to perform cancer-type classification.[78] To address high dimensionality and noise in genomic data, they selected a compact set of informative genes from thousands of candidates using seven feature-selection approaches, achieving high classification accuracies of 97.1% for leukemia and 93.6% for colon cancer. Fornstedt *et al.* utilized gradient boosting and support vector regression to predict retention times and peak widths for phosphorothioated oligonucleotides.[79] This model generates synthetic chromatograms and estimates peak resolution, allowing the automated optimization of separation conditions for therapeutic impurities. Even though these examples offer extremely high efficiency in exploring massive sequence spaces and optimizing complex experimental procedures, similar to machine learning for peptide models, their broad expansion remains dependent on the availability of high-quality training datasets, the understanding of "black-box" predictions, and the ability of models to

generalize across the inherent stochasticity of biological systems.

6. Application of generative design

Besides predicting the detailed hierarchical structures of biomolecules, such as proteins, and their interactions, machine learning algorithms can also be employed for inverse design, virtually predicting the chemical structures of biomacromolecules according to specific user requirements. Traditional virtual designs are predominantly based on the high-throughput screening of biomacromolecules.[82–91] In these works, the determinants of material properties, such as molecular structures, composition, reaction conditions, and processing parameters, are referred to as "genes". These genes encompass the chemical structure of repeating monomers, chain topology (including length, distribution, and sequence), and morphological features. By combining these genes according to established synthesis routes, the libraries of virtual candidates can be generated. Subsequently, machine learning models facilitate high-throughput property predictions, enabling the efficient screening and identification of promising polymers that meet targeted performance criteria. However, the efficacy of this reverse virtual design scheme largely relies on the quality of the virtual candidate library, which inherently limits the sampling space. Consequently, when designing a new structure, the process is restricted to selecting the best available candidate from the library rather than generating a novel model tailored to specific environmental constraints.[92,93]

Diffusion models offer a promising solution to address these limitations, enabling the conditional generation of chemical structures within specific microenvironments. RFdiffusion is a unified generative model for protein structures, which combines denoising and generative diffusion capabilities.[80] This allows it to design protein backbones and achieve state-of-the-art performance in both unconditional and topology-constrained protein monomer design, symmetric oligomer design, enzyme active site scaffolding, protein binder design and symmetric motif scaffolding for therapeutic and metal-binding protein design, starting from simple molecular blueprints. It is trained to reverse Gaussian noise to generate new protein structures *via* an iterative denoising process, as illustrated in Fig. 3(a). RFdiffusion has been fine-tuned with minimal architectural changes to integrate within the RoseTTAFold network, repurposed for generation. The main difference lies in the input: RoseTTAFold starts from a protein sequence, while RFdiffusion starts from diffused structural frames. Nonetheless, both models output final 3D atomic coordinates. Furthermore, the NCFlow model, rooted in the flow matching generative framework, provides an inherently systematic method to introduce noncanonical amino acids in the design of peptides,[81] as shown in Fig. 3(b). It is worth noting that this approach offers a level of conformation flexibility and chemical diversity that has never been considered before to carry out functional exploration in higher-dimensional space. The algorithm conducts high-throughput conformational generation over each residue position, candidate amino acid, and protein target. By combining learned and physics-based scoring functions, it accurately outputs structurally viable peptide variants, including the noncanonical amino acid(s), with improved binding affinity. For small-molecule binders, Ahern *et al.* construct RFdiffusion All-Atom (RFdiffusionAA) using a training of RFAA for the denoising task of protein structures that encase small molecules in diffusion.[94,95] RFdiffusionAA initializes small molecule–protein binding with random residue poses and uses diffusion to generate the protein structures that enclose the small molecule. The 2024 launch of AlphaFold3 marked a major shift in biomolecular modeling.[96,97] By replacing the Evoformer with a Pairformer module, the system cut computational overhead by 38% and reduced the dependency on multiple sequence alignments (MSAs). A main innovation lies in applying the diffusion-based module. It replaces traditional structural parameterization with a reverse denoising process to sample atomic coordinates directly. As a unified framework, it predicts interactions across proteins, nucleic acids, and small molecules simultaneously. Thus, while protein and small-molecule generation with machine learning has matured, as with the case of protein structure prediction, the inverse design of highly chemically engineered biomolecules has been difficult due to the unavailability of high-quality data. Additionally, even in well-studied tasks, such as protein structure generation, diffusion-based models suffer from high computational cost, slow sampling speed, and no fine-grained control.

7. Deep learning prediction of the ensemble properties of biomacromolecules

The utility of machine learning in studying self-assembly extends well beyond protein structures and interactions, encompassing the prediction of the critical ensemble properties of biomacromolecules, such as phase behavior and morphological outcomes. Wang Huaimin *et al.* have demonstrated a machine learning model (TransSAFP) that predicts the self-assembly of proteins through attention mechanisms and transfer learning techniques.[98] Its architecture is illustrated in Fig. 4(a). The model accurately describes the self-assembly behavior and biological function of the peptides with negligible experimental annotations in seconds, a process that would take humans billions of times longer. Leveraging this predictive power, they successfully designed novel SAFPs with potent antimicrobial activity, offering a promising solution to the growing challenge of bacterial drug resistance. Employing support vector machine (SVM) algorithms, Wang Huaimin *et al.* also successfully predicted aggregation propensity values for an exhaustive list of 160 000 tetrapeptide sequences, using training data from Martini molecular dynamics simulations.[100] Out of this pool, 55 peptides predicted by machine learning were synthesized chemically and tested to confirm their hydrogel-forming abilities, using an adapted APHC scoring function. With this modification, the prediction accuracy improved from 61.5% to 87.1%. Yang *et al.* introduced a novel data-driven

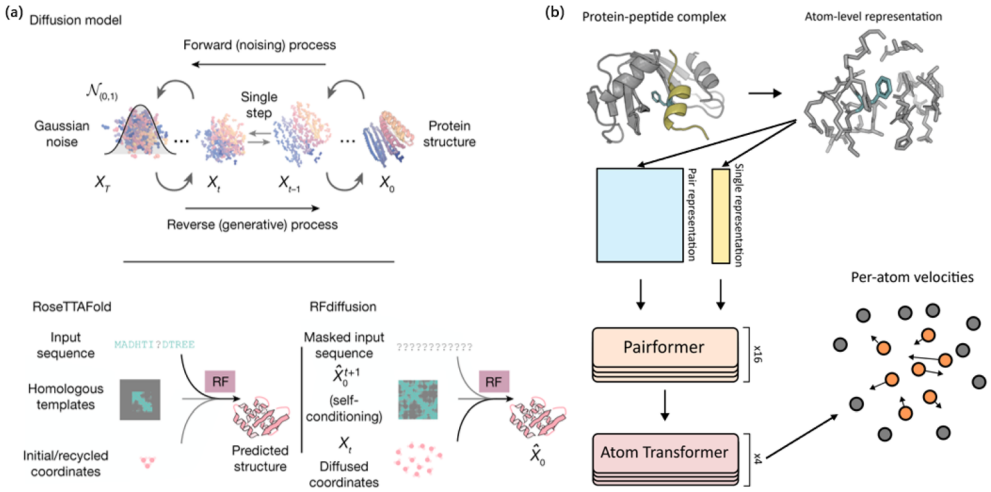

Fig. 3 Generative models. (a) RFdiffusion trained to recover the protein structure (XT) from Gaussian noise (X0). The architecture integrating this process with the RoseTTAFold network is illustrated in the lower panel (copied with permission,[80] copyright 2023, Joseph L. Watson *et al.*). (b) NCFlow model predicting the 3D conformation of a small molecule within a specified protein pocket, using the protein-peptide complex and the chemical graph of the small molecule (atom types and bond connectivity) as input (copied with permission[81]).

framework integrating machine learning (Random Forest and eXtreme Gradient Boost) and a custom deep neural network to accurately predict polysaccharide yield from cornstalks, achieving a peak accuracy of 95.6%.[101] By leveraging interpretable AI to quantify feature importance, identifying enzyme solution volume as the dominant factor, the approach effectively decodes complex enzymatic interactions. Abigail S. Knight *et al.* developed an integrated machine learning model to decipher sequence–conformation relationships in disordered synthetic polymers. Their strategy, depicted in Fig. 4(b), included a high-throughput workflow that integrates conformational characterization and secondary mass spectrometry sequencing, a color-classify image analysis model to quantify conformational propensities across libraries comprising more than 1000 polymers, and a new computational architecture called MotifFold, which combines a gradient-boosting regression algorithm with a frequency-embedding strategy.[99] Their work underscores the importance of discrete motifs governing the conformation of macromolecules. Biancalani *et al.* used GNEprop, which is based on graph neural networks (GNNs) and cross-attention mechanisms, to screen approximately 2 million small molecules against a sensitized *Escherichia coli* strain.[102] This method yields thousands of hits and reveals 82 structurally novel antibacterial compounds, providing a powerful method to

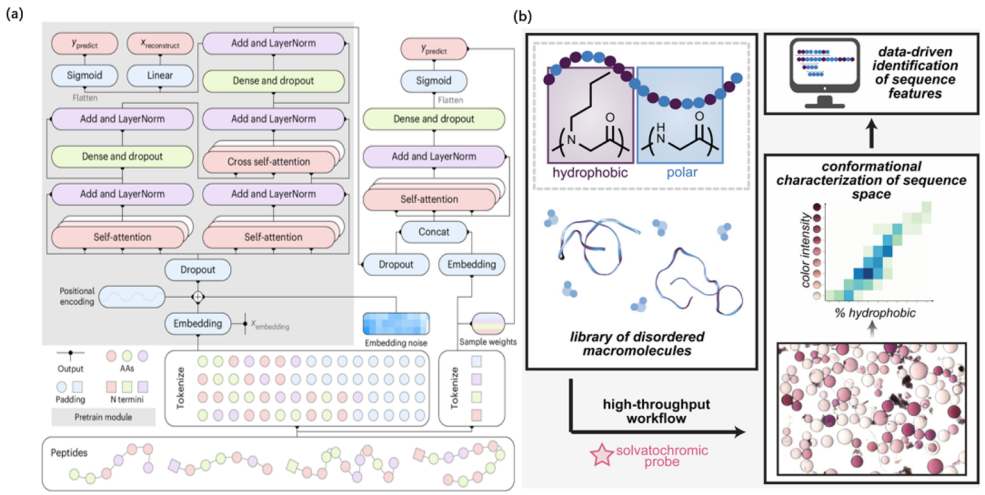

Fig. 4 Deep-learning prediction of ensemble properties. (a) TransSAFP model for self-assembling functional peptide discovery using a pretraining transfer learning architecture (copied with permission,[98] copyright 2025, the author(s), under exclusive licence to Springer Nature Limited). The pretraining module (left, grey) learns to predict antimicrobial activity and reconstruct peptide sequences from amino acid data. These learned representations are then transferred to a second module (right, white), which uses them to predict the activity of structurally altered peptides. (b) Schematic of the high-throughput workflow developed herein for analyzing a sequence space (copied with permission,[99] copyright 2024, Elsevier Inc.). The workflow integrates three steps: a one-bead one-compound library, a colorimetric assay, and motif analysis.

combat the growing antimicrobial resistance problem. A machine learning architecture proposed by Jiang *et al.* can predict protein circular dichroism (CD) spectra, nearly 10 thousand times faster than first-principles, which is in close agreement with experimental results.[103] The ab-initio-level accuracy achieved here is mainly due to employing embedded density descriptors and electric and magnetic transition dipole moments. Therefore, machine learning holds immense potential to revolutionize materials science by predicting sophisticated properties, ranging from self-assembly phase behavior to antimicrobial resistance efficiency, enabling the virtual screening of millions of candidate compounds. Additionally, from the application perspective, Beck *et al.* employed a pre-trained deep learning framework, the molecule transformer-drug target interaction (MT-DTI), to identify commercially available therapeutics with potential efficacy against SARS-CoV-2 viral proteins.[104] However, the practical application of machine learning in materials science is hampered by poor model generalization and a high dependency on training data. This often confines a model's application to its original context, permitting only minimal extrapolation. Consequently, developing a general-purpose materials AI remains particularly challenging.

8. Outlook and perspective

AI, particularly deep neural networks, is poised to profoundly reshape molecular design and biomolecular engineering; however, persistent challenges must be addressed to enable their widespread adoption and practical application. Modern AI methods, including attention architectures, deep generative models, and geometric deep learning, are highly effective for exploring highly complex biochemical spaces, not only for predicting structures but also for designing and discovering new sequences of functional peptides, proteins, and hybrid biomaterials. These methods have been effective at encoding the nuances of the sequence–structure–function correlations and even providing users with desired properties, significantly speeding up the design-test iterative process. However, generalization to highly engineered or non-biological molecular systems–such as chimeric peptide–polymer conjugates or highly modified biomacromolecules-is limited by paucity of data, as well as by the complexity of multi-modal interactions. An alternative method for dealing with current experimental limitations in biomolecular investigation is molecular dynamics (MD) simulation. A principal advantage of MD is its significantly lower cost compared with those of high-throughput experimental methods, enabling the efficient computation and quantitative analysis of specific atomic interactions.[105] Nonetheless, the validity of MD simulations against experimental benchmarks is not absolute but hinges on multiple aspects that warrant critical scrutiny: the selection of force field parameters, the configuration of the system under investigation, the extent to which environmental factors are incorporated, the utilization of optimization algorithms, and even the specific computational software used.[106–109] A model may yield highly accurate results for one scenario but prove inadequate for another. In turn, this implies that even though an MD simulation can elucidate the generic trends of a mechanism or observed qualitative behaviour, precise numerical values derived under specific conditions may not match the experimental results exactly. This intrinsic uncertainty requires the careful verification and an informed interpretation of computational results. Furthermore, many state-of-the-art AI models suffer from limited interpretability, high computational demand, and insufficient integration of thermodynamic or kinetic information, which constrain their potential to guide experimental synthesis and optimization. We believe that research in the future should focus on the design of extensible and physics-informed neural networks; the creation of open standards and multiscale biomolecular datasets; and the utilization of active learning and Bayesian optimization schemes for their generalizability, robustness, and experimental relevance. Additionally, apart from enhancing the capabilities of the diffusion model, we can also use VAEs in a complementary role to overcome some of its deficiencies. As is widely known, VAEs tend to converge more easily in training and are easier to interpret, due to their well-structured latent space. The diffusion model can be used in combination with a VAE model. An example of this is the latent diffusion model, which uses a VAE to compress images under a lower-dimensional latent space, where the diffusion process is applied for greater computational efficiency. By effectively combining the high-resolution fidelity and sampling efficiency of both diffusion and VAE models, Stable Diffusion has become one of the most widely used architectures.[110,111]

To facilitate the transition from conventional trial-and-error methodologies to data-driven discovery, a closed-loop inverse design framework can be employed to efficiently predict and engineer optimized chemical structures. Generative models, such as VAE and RFdiffusion, achieve this by encoding complex chemical parameters into a latent space, enabling the *de novo* synthesis of molecular architectures tailored to specific functional requirements. Apart from generative models, extensive dataset screening using forward predictive models could also be applied to predict the most plausible and efficient chemical structures based on the chemical parameters. They can be trained on sparse datasets, such as those utilizing attention mechanisms, to identify high-performance traits, such as structural stability. These models subsequently function as high-throughput screening frameworks, assigning scores to each constituent within extensive libraries to facilitate the selection of top-ranked candidates.[112] Additionally, given the vast sequence space of potential candidates, an active learning strategy in combination with computer simulation, such as molecular dynamics, can be employed to accelerate the screening process.[113] Initially, the target candidate is extracted from primary databases to establish an unlabeled candidate pool. An initial training set is then constructed by selecting samples, *via* random sampling or expert heuristics, and quantifying their performance through molecular dynamics simulations. Subsequently, an iterative optimization loop is initiated: the model predicts performance across the unlabeled dataset, while a query strategy designed to balance exploration

and exploitation identifies high-value candidates for simulation validation. By continuously updating the training set with these refined data points, the framework efficiently converges on peptide sequences with optimal encapsulation performance.

The AI predictive modeling is able to identify and mitigate potential adverse influences by assessing critical toxicological parameters in the early design phase. Leveraging advanced deep learning architectures, such as deep quantitative structure–activity relationship and graph neural networks, researchers can precisely screen candidate sequences for mutagenicity, carcinogenicity, and immunogenicity.[114,115] These models are capable of detecting "toxicophores", substructural motifs connected to adverse biological responses, contributing to a "safety-by-design" approach that filters high-risk candidates before resource-intensive experimental validation. Even though their efficacy is greatly increased through processing high-dimensional datasets, their primary advantage lies in reducing experimental cost.

While considerable progress has been achieved, a key constraint remains in the heterogeneity, incompleteness, and experimental uncertainty of available datasets, which impedes the reliable and systematic training of AI models. Different AI models or research topics require different degrees of supervision; some demand high-quality annotated data for fully supervised learning, while others rely on sparse, noisy datasets under semi-supervised conditions. Nevertheless, many reported studies often neglect how such differences in the dataset should be considered when selecting model architectures. These limitations in data quality could potentially lead to a reproducibility gap, where AI models trained on particular experimental datasets cannot reliably perform across different experimental conditions or measurement setups. In the future, research efforts should be made to establish standardized benchmark systems that directly measure and account for data uncertainty and the scope of application. This could contribute to ensuring that AI model predictions yield consistent, reproducible findings with genuine practical relevance for scientific investigation.

The tension between statistical pattern recognition and physical mechanism comprises a fundamental challenge in biomolecular design, where the success of "black-box" predictions often comes at the expense of physical validity. While deep learning architectures demonstrate exceptional performance in identifying high-dimensional correlations within large-scale datasets, the representations they learn frequently circumvent the fundamental laws of thermodynamics, leading to predictions that may be statistically feasible yet physically unreasonable. To bridge this critical gap, explainability must transit from simple post-hoc visualizations toward the robust integration of inductive biases and hybrid methodologies. Architectures such as equivariant neural networks and physics-informed loss functions are essential for ensuring that model outputs adhere to physical and chemical constraints. What's more, incorporating explainable artificial intelligence frameworks, including SHapley Additive exPlanations, MolAnchor, or counterfactual explanations, can provide necessary transparency into the model's decision-making process.[116] Beyond purely algorithmic adjustments, the synergy between molecular dynamics simulations and deep learning offers a powerful solution. Molecular dynamics can either be coupled with AI models to enforce chemical constraints or used to generate physically meaningful parameters that serve as enriched inputs. Ultimately, the credibility of modern generative methods in biomolecular design relies on balancing flexibility with established chemical rules. By anchoring statistical learning in physical realism, we produce designs that are both plausible and viable, turning models from pattern matchers into reliable, physically grounded tools for scientific advancement.

LLMs, positioned at the forefront of artificial general intelligence (AGI) research due to their capacity to process and generate human language and complex patterns, represent a transformative force with the potential to fundamentally reshape the paradigms of molecular design and biomolecular engineering.[117,118] The use of these models as part of this scientific workflow represents an important step toward the future. The evolution of standard, "tool by tool", computational practice toward a richer, more intuitive and semantically conscious design process is underway. In this regard, LLMs can take the role of intelligent interfaces that interpret, say, the research objective in natural language, for example, "design a cyclic peptide that inhibits protein X with enhanced oral bioavailability", and map such instructions to a sequence of computational operations that would achieve that objective. It spans the entire pipeline: mining literature for synthesis candidates, automating compound generation, executing custom simulations, and distilling complex technical results as readable documents. Thus, LLMs enable non-coding researchers to harness sophisticated AI *via* natural language, bridging the gap between experimentalists and computational design. This promises a future of collaborative AI partners capable of multi-scale reasoning to drive innovation across scientific disciplines.

9. Conclusion

AI is fundamentally reshaping and transforming the development of biomacromolecule engineering. The transition from predictive and descriptive technologies, exemplified by AlphaFold, to generative platforms, such as RFdiffusion, marks a significant evolution in the field. This transition enables a rational, data-driven approach for predicting protein structure, engineering protein–drug interactions, and generating novel functional structures. Furthermore, AI tools also enable the prediction of self-assembly behavior, the design of high-affinity binders, and the establishment of novel structure–function relationships. By accelerating the design-build-test cycle, these AI-driven methods are unlocking unprecedented possibilities, heralding a new era of rapid innovation in biotechnology and materials science.

Author contributions

Hairui Li: writing – review and editing; Guocheng Zhang: conceptualization, supervision, project administration, and

writing – review and editing; Yang Liu: conceptualization, supervision, project administration, and writing – review and editing; Bin Yan: writing – review and editing; Zaihang Ye: writing – review and editing; and Xudong Zhen: writing – review and editing.

Conflicts of interest

There are no conflicts to declare.

Data availability

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

Acknowledgements

This work was financially supported by the National Natural Science Foundation of China (52303298 and 52205582), the Application and Basic Research of Sichuan Department of Science and Technology, the Project of State Key Laboratory of Polymer Materials Engineering, the Natural Science Foundation of Henan Province (252300421321), and the Key Scientific Research Projects of Universities in Henan (22B460001). During the preparation of this work, the authors used DeepSeek-R1 to improve the language readability only. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

1 A. Martinez-Rumayor, A. M. Richards, J. C. Burnett and J. L. Januzzi Jr, *Am. J. Cardiol.*, 2008, **101**, S3–S8.
2 D. Vaudry, A. Falluel-Morel, S. Bourgault, M. Basille, D. Burel, O. Wurtz, A. Fournier, B. K. C. Chow, H. Hashimoto, L. Galas and H. Vaudry, *Pharmacol. Rev.*, 2000, **52**, 269–324.
3 Q. Wang, Y. N. Fu, S. J. Sun, C. Y. Huang, Y. F. Yi, J. Q. Wang, Y. Deng and M. Y. Wu, *Chin. Chem. Lett.*, 2023, **34**, 107508.
4 D. Bhandari, S. Rafiq, Y. Gat, P. Gat, R. Waghmare and V. Kumar, *Int. J. Pept. Res. Ther.*, 2020, **26**, 139–150.
5 S. N. Liu, J. H. Meng, L. Y. Cui, H. Chen, L. Q. Shi and R. J. Ma, *Chin. J. Polym. Sci.*, 2024, **42**, 559–569.
6 P. E. Wright and H. J. Dyson, *J. Mol. Biol.*, 1999, **293**, 321–331.
7 F. L. Zhang and P. J. Casey, *Annu. Rev. Biochem.*, 1996, **65**, 241–269.
8 Y. Z. Zheng, S. Sun, J. L. Liu, Q. Y. Zhao, H. Zhang, J. Zhang, P. Zhou, Z. K. Xiong, C. S. He and B. Lai, *Chin. Chem. Lett.*, 2025, **36**, 110722.
9 D. Pandey, K. Niwaria and B. Chourasia, *Mach. Learn.*, 2019, **6**, 916–922.
10 S. B. Kotsiantis, I. D. Zaharakis and P. E. Pintelas, *Artif. Intell. Rev.*, 2006, **26**, 159–190.
11 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, *Adv. Neural Inf. Process. Syst.*, 2017, **30**, 6000–6010.
12 A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan and D. S. W. Ting, *Nat. Med.*, 2023, **29**, 1930–1940.
13 D. Bank, N. Koenigstein, R. Giryes, *arXiv*, 2020, preprint, arXiv:2003.05991, DOI: 10.48550/arXiv.2003.05991.
14 Z. Xing, Q. Feng, H. Chen, Q. Dai, H. Hu, H. Xu, Z. Wu and Y. Jiang, *ACM Comput. Surv.*, 2023, **57**, 1–42.
15 M. Schuster and K. K. Paliwal, *IEEE Trans. Signal Process.*, 1997, **45**, 2673–2681.
16 Z. F. Zheng, Y. Z. H. Wang, Y. X. Huang, S. C. Song, M. C. Yang, B. Tang, F. Y. Xiong and Z. Y. Li, *Patterns*, 2025, **6**, 2101176.
17 A. B. Artzy, R. Schwartz, *Presented in Part at Conference 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, Miami, November, 2024.
18 R. Patil and V. Gudivada, *Appl. Sci.*, 2024, **14**, 2074.
19 S. N. Prasad Kumar, R. Gangurde, U. L. Mohite, *et al.*, *Int. J. Comput. Intell*. Appl., 2025.
20 J. Kaplan, S. McCandlish, T. Henighan, T. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, *arXiv*, 2020, preprint, arXiv:2001.08361, DOI: 10.48550/arXiv.2001.08361.
21 J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. Casas, J. Welbl, W. Saunders, A. Gelman, A. Clark, C. Berner, F. Petroni, M. Henderson, R. Ring, E. Young, S. Riedel, A. Botev, A. Balwit, S. Misra, *arXiv*, 2022, preprint, arXiv:2203.15556, DOI: 10.48550/arXiv.2203.15556.
22 J. Wei, R. B. Y. Tay, C. Raffel, B. Zoph, S. Borgeaud, W. Fedus, Y. Liu, S. Narang, T. Salimans, D. Schuurmans, X. Shi, Y. Tsvetkov, N. Welleck, *arXiv*, 2022, preprint, arXiv:2206.07682, DOI: 10.48550/arXiv.2206.07682.
23 T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, M. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei, *Adv. Neural Inf. Process. Syst.*, 2020, **33**, 1877–1901.
24 J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le, *arXiv*, 2022, preprint, arXiv:2109.01652, DOI: 10.48550/arXiv.2109.01652.
25 H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, Y. Xu, M. Shoeybi, M. Patwary, R. Puri, P. Fung, A. Anandkumar, B. Catanzaro and M. Lewis, *J. Mach. Learn. Res.*, 2024, **25**, 53.
26 J. Wei, X. Z. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le and D. Zhou, *Adv. Neural Inf. Process. Syst.*, 2022, **35**, 24824–24837.
27 C. Zhou, Q. Li, C. Li, J. Yu, Y. X. Liu, G. J. Wang, K. Zhang, C. Ji, Q. B. Yan, L. F. He, H. Peng, J. X. Li, J. Wu, Z. W. Liu, P. T. Xie, C. M. Xiong, J. Pei, P. S. Yu and L. C. Sun, *Int. J. Mach. Learn. Cybern.*, 2023, 1–65.

28 A. Chowdhery, S. Narang, J. Devlin, M. Bosma, S. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, Y. Tsvetkov, M. Bos, D. Zhou, D. Metzler, E. H. Chi, P. Crook, J. Dean, M. I. Petrov, W. Fedus, M. Patwary, R. Puri, M. Shoeybi, B. Catanzaro, W. Fedus, A. Anandkumar, A. Y. Ng, Q. V. Le, I. Mordatch and I. Sutskever, *J. Mach. Learn. Res.*, 2023, **24**, 113.

29 D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, G. Irving, *arXiv*, 2022, preprint, arXiv:1909.08593, DOI: 10.48550/arXiv.1909.08593.

30 P. F. Christiano, J. Leike, T. B. Brown, J. Martens and S. Legg, *Adv. Neural Inf. Process. Syst.*, 2017, **30**, 4302–4310.

31 J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, *arXiv*, 2017, preprint, arXiv:1707.06347, DOI: 10.48550/arXiv.1707.06347.

32 L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, J. Martens, I. Sutskever, *arXiv*, 2022, preprint, arXiv:2203.02155, DOI: 10.48550/arXiv.2203.02155.

33 Z. Li, M. J. Jiang, S. Wang and S. G. Zhang, *Drug Discov. Today*, 2022, **27**, 103373.

34 A. Y. T. Wang, S. K. Kauwe, R. J. Murdock, *et al.*, *npj Comput. Mater.*, 2021, **7**, 77.

35 A. Sultan, J. Sieg, M. Mathea and A. Volkamer, *J. Chem. Inf. Model.*, 2024, **64**, 6259–6280.

36 C. Doersch, *arXiv*, 2016, preprint, arXiv:1606.05908, DOI: 10.48550/arXiv.1606.05908.

37 L. P. Cinelli, M. A. Marins, E. A. B. d. Silva, S. L. Netto, *Variational Methods for Machine Learning with Applications to Deep Networks*, Springer, 2021.

38 J. Ho, A. Jain and P. Abbeel, *Adv. Neural Inf. Process. Syst.*, 2020, **33**, 6840–6851.

39 B. R. Beck, B. Shin, Y. Choi, S. Park and K. S. Kang, *Comput. Struct. Biotechnol. J.*, 2020, **18**, 784–790.

40 Y. Song and S. Ermon, *Adv. Neural Inf. Process. Syst.*, 2019, **32**, 11886–11898.

41 Y. Song and S. Ermon, *Adv. Neural Inf. Process. Syst.*, 2020, **33**, 12438–12448.

42 T. Karras, M. Aittala, T. Aila, J. Hellsten, J. Lehtinen and S. Laine, *Adv. Neural Inf. Process. Syst.*, 2022, 26565–26577.

43 Y. Song, C. Durkan, I. Murray and S. Ermon, *Adv. Neural Inf. Process. Syst.*, 2021, **34**, 1415–1428.

44 J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, *Nature*, 2021, **596**, 583–589.

45 M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. S. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. P. Rodrigues, A. A. van Dijk, A. Ebrecht, U. Oppermann, A. Rives, S. Velankar, F. Xia, X. Zuo and D. Baker, *Science*, 2021, **373**, 871–876.

46 A. Chatr-Aryamontri, B.-J. Breitkreutz, R. Oughtred, L. Boucher, S. Heinicke, D. Chen, C. Stark, A. Breitkreutz, N. Kolas, L. O'Donnell, T. Reguly, J. Nixon, L. Ramage, A. Winter, A. Sellam, C. Chang, J. Hirschman, C. Theesfeld, J. Rust, M. S. Livstone, K. Dolinski and M. Tyers, *Nucleic Acids Res.*, 2015, **43**, D470–D478.

47 A. Chatr-Aryamontri, R. Oughtred, L. Boucher, J. Rust, C. Chang, N. K. Kolas, L. O'Donnell, S. Oster, C. Theesfeld, A. Sellam, C. Stark, B.-J. Breitkreutz, K. Dolinski and M. Tyers, *Nucleic Acids Res.*, 2017, **45**, D369–D379.

48 R. Oughtred, C. Stark, B.-J. Breitkreutz, J. Rust, L. Boucher, C. Chang, N. Kolas, L. O'Donnell, G. Leung, R. McAdam, F. Zhang, S. Dolma, A. Willems, J. Coulombe-Huntington, A. Chatr-Aryamontri, K. Dolinski and M. Tyers, *Nucleic Acids Res.*, 2019, **47**, D529–D541.

49 J. D. Peri, R. Navarro, T. Z. Amanchy, C. K. Kristiansen, V. Jonnalagadda, V. Surendranath, B. Niranjan, T. K. B. Muthusamy, M. Gandhi, N. Gronborg, N. Ibarrola, K. Deshpande, H. N. Shanker, B. P. Shivashankar, M. A. Rashmi, Z. X. Ramya, K. N. Zhao, N. Chandrika, H. C. Padma, A. J. Harsha, M. P. Yatish, M. Kavitha, D. R. Menezes, S. Choudhury, N. Suresh, R. Ghosh, S. Saravana, S. Chandran, M. Krishna, *et al.*, *Genome Res.*, 2003, **13**, 2363.

50 J. Kypr, I. Kejnovská, D. Renciuk and M. Vorlícková, *Nucleic Acids Res.*, 2009, **37**, 1713–1725.

51 G. R. Mishra, M. Suresh, K. Kumaran, N. Kannabiran, S. Suresh, P. Bala, K. Shivakumar, N. Anuradha, R. Reddy, T. M. Raghavan, S. Menon, G. Hanumanthu, M. Gupta, S. Upendran, S. Gupta, M. Mahesh, B. Jacob, P. Mathew, P. Chatterjee, K. S. Arun, S. Sharma, K. N. Chandrika, N. Deshpande, K. Palvankar, R. Raghavnath, R. Krishnakanth, H. Karathia, B. Rekha, R. Nayak, G. Vishnupriya, H. G. M. Kumar, M. Nagini, G. S. S. Kumar, R. Jose, P. Deepthi, S. S. Mohan, T. K. B. Gandhi, H. C. Harsha, K. S. Deshpande, M. Sarker, T. S. K. Prasad and A. Pandey, *Nucleic Acids Res.*, 2006, **34**, D411–D414.

52 S. Peri, J. D. Navarro, R. Amanchy, T. Z. Kristiansen, C. K. Jonnalagadda, V. Surendranath, V. Niranjan, B. Muthusamy, T. K. B. Gandhi, M. Gronborg, N. Ibarrola, N. Deshpande, K. Shanker, H. N. Shivashankar, B. P. Rashmi, M. A. Ramya, Z. Zhao, K. N. Chandrika, N. Padma, H. C. Harsha, A. J. Yatish, M. P. Kavitha, M. Menezes, D. R. Choudhury, S. Suresh, N. Ghosh, R. Saravana, S. Chandran, S. Krishna, M. Joy, S. K. Anand, V. Madavan, A. Joseph, G. W. Wong, W. P. Schiemann, S. N. Constantinescu, L. Huang, R. Khosravi-Far, H. Steen, M. Tewari, S. Ghaffari, G. C. Blobe, C. V. Dang, J. G. N. Garcia, J. Pevsner, O. N. Jensen, P. Roepstorff, K. S. Deshpande, A. M. Chinnaiyan, A. Hamosh,

A. Chakravarti and A. Pandey, *Genome Res.*, 2003, **13**, 2363–2371.
53 R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi and L.-S. L. Yeh, *Nucleic Acids Res.*, 2004, **32**, D115–D119.
54 U. P. Consortium, *Nucleic Acids Res.*, 2017, **45**, D158–D169.
55 U. P. Consortium, *Nucleic Acids Res.*, 2023, **51**, D523–D531.
56 U. P. Consortium, *Nucleic Acids Res.*, 2018, **46**, 2699.
57 L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie and D. Eisenberg, *Nucleic Acids Res.*, 2004, **32**, D449–D451.
58 I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte and D. Eisenberg, *Nucleic Acids Res.*, 2000, **28**, 289–291.
59 I. Xenarios, L. Salwínski, X. J. Duan, P. Higney, S.-M. Kim and D. Eisenberg, *Nucleic Acids Res.*, 2002, **30**, 303–305.
60 D. Szklarczyk, A. L. Gable, K. C. Nastou, D. Lyon, R. Kirsch, S. Pyysalo, N. T. Doncheva, M. Legeay, T. Fang, P. Bork, L. J. Jensen and C. von Mering, *Nucleic Acids Res.*, 2021, **49**, D605–D612.
61 D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen and C. von Mering, *Nucleic Acids Res.*, 2019, **47**, D607–D613.
62 X.-W. Chen and M. Liu, *Bioinformatics*, 2005, **21**, 4394–4400.
63 J.-F. Xia, K. Han and D.-S. Huang, *Protein Pept. Lett.*, 2010, **17**, 137–145.
64 S. Hirose, K. Yokota, Y. Kuroda, H. Wako, S. Endo, S. Kanai and T. Noguchi, *BMC Struct Biol.*, 2010, **10**, 20.
65 U. K. Muppirala, V. G. Honavar and D. Dobbs, *BMC Bioinf.*, 2011, **12**, 489.
66 G. Taherzadeh, Y. Zhou, A. W.-C. Liew and Y. Yang, *Bioinformatics*, 2018, **34**, 477–484.
67 S. Dohkan, A. Koike and T. Takagi, *In Silico Biol.*, 2006, **6**, 515–529.
68 J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li and H. Jiang, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 4337–4341.
69 C. W. Cheng, E. C. Su, J. Hwang, T. Y. Sung and W. L. Hsu, *BMC Bioinf.*, 2008, **9**, S6.
70 Y. Guo, L. Yu, Z. Wen and M. Li, *Nucleic Acids Res.*, 2008, **36**, 3025–3030.
71 M.-G. Shi, J.-F. Xia, X.-L. Li and D.-S. Huang, *Amino Acids*, 2010, **38**, 891–899.
72 L. T. Chen, Z. Quinn, M. Dumas, C. Peng, L. Hong, M. Lopez-Gonzalez, A. Mestre, R. Watson, S. Vincoff, L. Zhao, J. L. Wu, A. Stavrand, M. Schaepers-Cheu, T. Z. Wang, D. Srijay, C. Monticello, P. Vure, R. Pulugurta, S. Pertsemlidis, K. Kholina, S. Goel, M. P. DeLisa, J. T. A. Chi, R. Truant, H. C. Aguilar and P. Chatterjee, *Nat. Biotechnol.*, 2025, DOI: 10.1038/s41587-025-02761-2.
73 C. Kuenneth and R. Ramprasad, *Nat. Commun.*, 2023, **14**, 4099.
74 H. Qiu and Z. Y. Sun, *Npj Comput. Mater.*, 2024, **10**, 273.
75 H. Qiu, L. Liu, X. Qiu, X. Dai, X. Ji and Z.-Y. Sun, *Chem. Sci.*, 2024, **15**, 534–544.
76 A. Lamens and J. Bajorath, *Chem. Sci.*, 2026, **17**, 1411–1422.
77 J. Ferraz-Caetano, F. Teixeira, M. N. D. S. Cordeiro and T. Miyao, *J. Catal.*, 2026, **453**, 116537.
78 S.-B. Cho and H.-H. Won, *Asia-Pac. Bioinf. Conf.*, 2003, **19**, 189–198.
79 J. Samuelsson, M. Enmark, G. Szabados, M. Rahal, B. S. Ahmed, J. Häggström, P. Forssén and T. Fornstedt, *J. Chromatogr. A*, 2025, **1747**, 465746.
80 J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. Vázquez Torres, A. Lauko, V. De Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek and D. Baker, *Nature*, 2023, **620**, 1089–1100.
81 J. S. Lee, P. M. Kim, *bioRxiv*, 2025, preprint.
82 S. Zhang, S. Du, L. Wang, J. Lin, L. Du, X. Xu and L. Gao, *Chem. Eng. J.*, 2022, **448**, 137643.
83 J. Zhu, M. Chu, Z. Chen, L. Wang, J. Lin and L. Du, *Chem. Mater.*, 2020, **32**, 4527–4535.
84 A. Mannodi-Kanakkithodi, G. Pilania, R. Ramprasad, T. Lookman and J. E. Gubernatis, *Comput. Mater. Sci.*, 2016, **125**, 92–99.
85 L. Tao, J. He, N. E. Munyaneza, V. Varshney, W. Chen, G. Liu and Y. Li, *Chem. Eng. J.*, 2023, **465**, 142949.
86 H. Qiu, X. Qiu, X. Dai and Z.-Y. Sun, *J. Mater. Chem. C*, 2023, **11**, 2930–2940.
87 J. M. Dean, N. E. Verghese, H. Q. Pham and F. S. Bates, *Macromolecules*, 2003, **36**, 9267–9270.
88 X. Xu, W. Zhao, L. Wang, J. Lin and L. Du, *Chem. Sci.*, 2023, **14**, 10203–10211.
89 R. van de Schoot, S. Depaoli, R. King, B. Kramer, K. Märtens, M. G. Tadesse, M. Vannucci, A. Gelman, D. Veen, J. Willemsen and C. Yau, *Nat. Rev. Methods Primers*, 2021, **1**, 1.
90 S. Zhao, T. Cai, L. Zhang, W. Li and J. Lin, *ACS Macro Lett.*, 2021, **10**, 598–602.
91 R. Gurnani, D. Kamal, H. Tran, H. Sahu, K. Scharm, U. Ashraf and R. Ramprasad, *Chem. Mater.*, 2021, **33**, 7008–7016.
92 M. J. Buehler, *Acc. Chem. Res.*, 2022, **55**, 3387–3403.
93 E. Vargo, J. C. Dahl, K. M. Evans, T. Khan, P. Alivisatos and T. Xu, *Adv. Mater.*, 2022, **34**, 2203168.
94 W. Ahern, PhD Thesis, University of Washington, 2025.
95 W. Ahern, J. Yim, D. Tischer, S. Salike, S. M. Woodbury, D. Kim, I. Kalvet, Y. Kipnis, B. Coventry, H. R. Altae-Tran, M. Bauer, R. Barzilay, T. S. Jaakkola, R. Krishna, D. Baker, *bioRxiv*, 2025, preprint.
96 J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong,

S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis and J. M. Jumper, *Nature*, 2024, **630**, 493–500.
97 D. Desai, S. V. Kantliwala, J. Vybhavi, R. Ravi, H. Patel and J. Patel, *Cureus*, 2024, **16**(7), e63646.
98 H. Y. Liu, Z. L. Song, Y. Zhang, B. H. Wu, D. H. Chen, Z. Zhou, H. Y. Zhang, S. S. Li, X. P. Feng, J. Huang and H. M. Wang, *Nat. Mater.*, 2025, **24**, 1295–1306.
99 E. C. Day, S. S. Chittari, K. C. Cunha, R. J. Zhao, J. N. Dodds, D. C. Davis, E. S. Baker, R. B. Berlow, J. E. Shea, R. U. Kulkarni and A. S. Knight, *Chem*, 2024, **10**, 3444–3458.
100 T. Y. Xu, J. Q. Wang, S. Zhao, D. H. Chen, H. Y. Zhang, Y. Fang, N. Kong, Z. Zhou, W. B. Li and H. M. Wang, *Nat. Commun.*, 2023, **14**, 3880.
101 Y. Tian, X. Yang, N. Chen, C. Li and W. Yang, *Environ. Sci. Ecotechnol.*, 2024, **19**, 100321.
102 G. Scalia, S. T. Rutherford, Z. Q. Lu, K. R. Buchholz, N. Skelton, K. Chuang, N. Diamant, J.-C. Hütter, J.-M. Luescher, A. Miu, J. Blaney, L. Gendelev, E. Skippington, G. Zynda, N. Dickson, M. Koziarski, Y. Bengio, A. Regev, M.-W. Tan and T. Biancalani, *Nat. Biotechnol.*, 2025, 1–14.
103 L. Y. Zhao, J. X. Zhang, Y. L. Zhang, S. Ye, G. Z. Zhang, X. Chen, B. Jiang and J. Jiang, *Jacs Au*, 2021, **1**, 2377–2384.
104 B. R. Beck, B. Shin, Y. Choi, S. Park and K. Kang, *Comput. Struct. Biotechnol. J.*, 2020, **18**, 784–790.
105 A. Wang, W. K. Liu, X. H. Jin, H. C. Wu, D. F. Zhang, X. L. Han, Y. Liu, Z. Li, M. M. Ding, J. H. Li and H. Tan, *Nano Lett.*, 2025, **25**, 7560–7567.
106 H. J. C. Berendsen, *Presented in Part at Conference the 2nd International Symposium on Algorithms for Macromolecular Modelling*, Berlin, May, 1997.
107 S. Jo, T. Kim, V. G. Iyer and W. Im, *J. Comput. Chem.*, 2008, **29**, 1859–1865.
108 J. Lee, X. Cheng, S. Jo, A. D. MacKerell, J. B. Klauda and W. Im, *J. Chem. Theory Comput.*, 2016, **12**, 405–413.
109 Y. Z. Lin, L. K. Feng, Y. D. Li, C. F. Chang, C. Z. Zhu, M. L. Wang and J. Xu, *Chin. J. Polym. Sci.*, 2024, **42**, 655–662.
110 R. Tang, L. Q. Liu, A. Pandey, Z. Y. Jiang, G. F. Yang, K. Kumar, P. Stenetorp, J. Lin, F. Ture, *arXiv*, 2022, preprint, arXiv:2210.04885, DOI: 10.48550/arXiv.2210.04885.
111 D. Bolya, J. Hoffman, *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, 4599–4603.
112 M. Liu, Y. Zhou, X. Mei, Z. Yu, B. Guan, Y. Xiao, S. Liu, H. Wang and Y. Qin, *Front. Cell Dev. Biol.*, 2025, **13**, 1755565.
113 A. Krishnakumar. *Active Learning Literature survey[J]. Tech. Rep.*, University of California, Santa Cruz, 2007, 42.
114 S. Li, L. Zhang, H. Feng, J. Meng, D. Xie, L. Yi, I. T. Arkin and H. Liu, *Interdiscip. Sci. Comput. Life Sci.*, 2021, **13**, 25–33.
115 C. Hung and G. Gini, *Mol. Divers.*, 2021, **25**, 1283–1299.
116 A. Lamens and J. Bajorath, *Chem. Sci.*, 2025, **17**, 1411–1422.
117 A. F. Mohammad, B. Clark, R. Agarwal, S. Summers and I. E. E. E. Cong, *Comput. Sci. Comput. Eng. Appl. Comput.*, 2023, 413.
118 D. Ilić and G. E. Gignac, *Intelligence*, 2024, **106**, 101858.