

# A comprehensive survey on artificial intelligence for biomolecule design

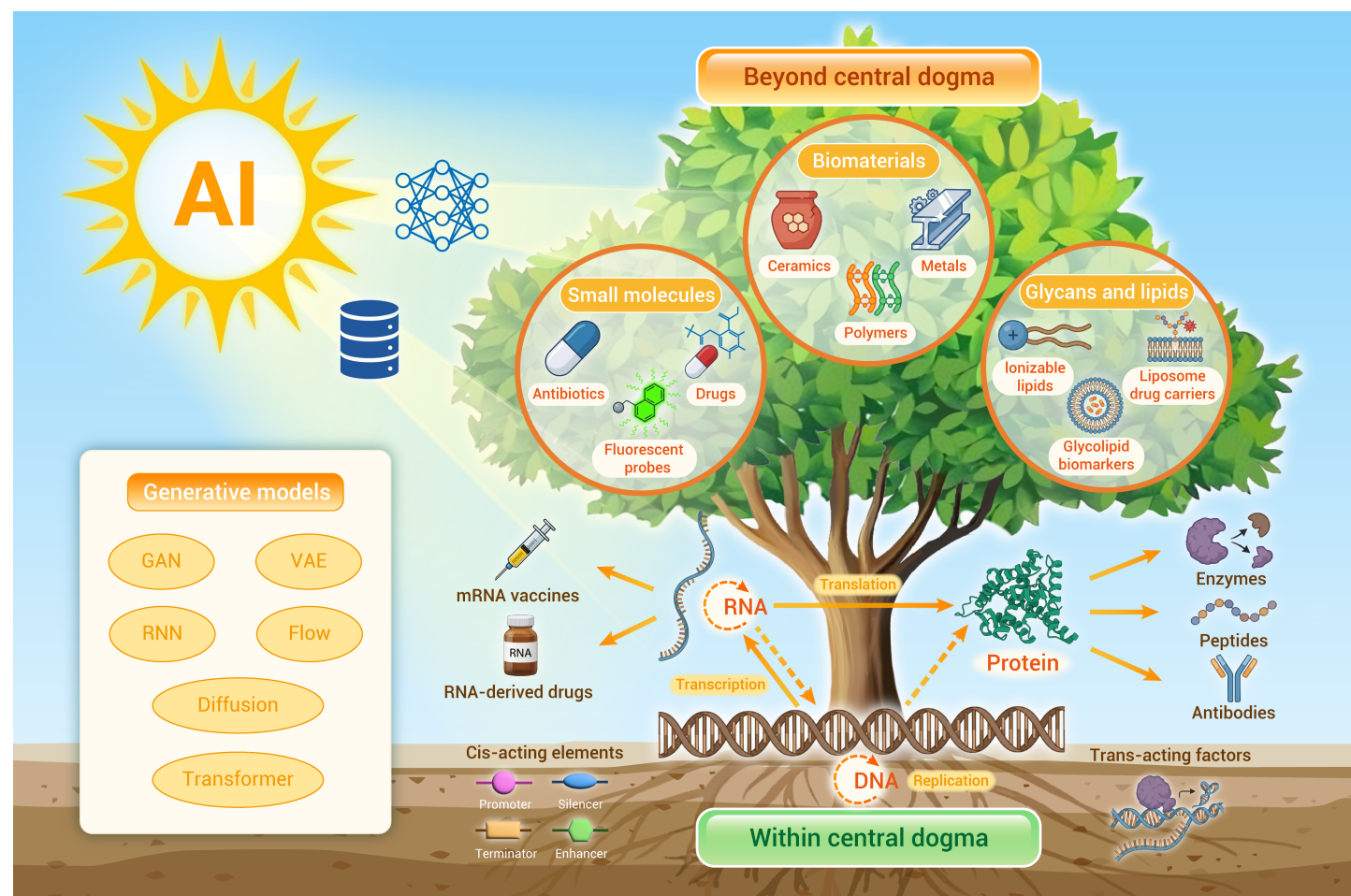
Guang Yang,<sup>1</sup> Jianing Li,<sup>2</sup> Qiong Wang,<sup>1</sup> Lianlian Wu,<sup>3,4</sup> Zhengshan Chen,<sup>3</sup> Peng-Cheng Zhao,<sup>1</sup> Haoyang Wang,<sup>1</sup> Qinglong Wang,<sup>1</sup> Bing Wen,<sup>1</sup> Chao Song,<sup>2</sup> Wenxin Xu,<sup>2</sup> Xiongfei He,<sup>1</sup> Xiao Zhang,<sup>2</sup> Sophia Tsoka,<sup>5,\*</sup> Siu-Ming Yiu,<sup>6,\*</sup> Fang-Xiang Wu,<sup>7,\*</sup> Meng Xiao,<sup>8,9,\*</sup> Shirui Pan,<sup>10,\*</sup> Hui Yu,<sup>2,\*</sup> Song He,<sup>3,\*</sup> Xiaoli Li,<sup>11,\*</sup> Xiaochen Bo,<sup>3,\*</sup> Min Wu,<sup>12,\*</sup> and Jian-Yu Shi<sup>1,\*</sup>

\*Correspondence: sophia.tsoka@kcl.ac.uk (S.T.); smyiu@cs.hku.hk (S.M.Y.); fangxiang.wu@usask.ca (F.X.W.); shaow@cnic.cn (M.X.); s.pan@griffith.edu.au (S.P.); huiyu@nwpu.edu.cn (H.Y.); hes1224@163.com (S.H.); xiaoli\_li@sutd.edu.sg (X.L.); boxiaoc@163.com (X.B.); wumin@a-star.edu.sg (M.W.); jianyushi@nwpu.edu.cn (J.Y.S.)

Received: April 29, 2025; Accepted: January 9, 2026; Published Online: January 17, 2026; <https://doi.org/10.59717/j.xinn-life.2026.100196>

© 2026 The Author(s). This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>).

## GRAPHICAL ABSTRACT



## PUBLIC SUMMARY

- Generative AI is reshaping molecule design across drug discovery, materials science, and synthetic biology.
- The central dogma framework organizes AI methods for DNA, RNA, proteins, drug-like small molecules, glycans, lipids, and biomaterials.
- Key advances, challenges, and future directions in AI-driven molecular engineering are systematically reviewed.

# A comprehensive survey on artificial intelligence for biomolecule design

Guang Yang,<sup>1</sup> Jianing Li,<sup>2</sup> Qiong Wang,<sup>1</sup> Lianlian Wu,<sup>3,4</sup> Zhengshan Chen,<sup>3</sup> Peng-Cheng Zhao,<sup>1</sup> Haoyang Wang,<sup>1</sup> Qinglong Wang,<sup>1</sup> Bing Wen,<sup>1</sup> Chao Song,<sup>2</sup> Wenxin Xu,<sup>2</sup> Xiongfei He,<sup>1</sup> Xiao Zhang,<sup>2</sup> Sophia Tsoka,<sup>5,\*</sup> Siu-Ming Yiu,<sup>6,\*</sup> Fang-Xiang Wu,<sup>7,\*</sup> Meng Xiao,<sup>8,9,\*</sup> Shirui Pan,<sup>10,\*</sup> Hui Yu,<sup>2,\*</sup> Song He,<sup>3,\*</sup> Xiaoli Li,<sup>11,\*</sup> Xiaochen Bo,<sup>3,\*</sup> Min Wu,<sup>12,\*</sup> and Jian-Yu Shi<sup>1,\*</sup>

<sup>1</sup>School of Life Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China

<sup>2</sup>School of Computer Science, Northwestern Polytechnical University, Xi'an 710072 China

<sup>3</sup>Department of Advanced & Interdisciplinary Biotechnology, Academy of Military Medical Sciences, Beijing 100850, China

<sup>4</sup>Academy of Medical Engineering and Translational Medicine, Tianjin University, Tianjin 300072, China

<sup>5</sup>Department of Informatics, King's College London, London WC2R 2LS, United Kingdom

<sup>6</sup>Department of Computer Science, The University of Hong Kong, Hong Kong 999077, China

<sup>7</sup>Division of Biomedical Engineering, University of Saskatchewan, Saskatoon S7N 5A2, Canada

<sup>8</sup>Computer Network Information Center, Chinese Academy of Sciences, Beijing 100864, China

<sup>9</sup>Duke-NUS Medical School, National University of Singapore, Singapore 169857, Singapore

<sup>10</sup>School of Information and Communication Technology, Griffith University, Gold Coast QLD 4222, Australia

<sup>11</sup>Information Systems Technology and Design, Singapore University of Technology and Design, Singapore 487372, Singapore

<sup>12</sup>Institute for Infocomm Research, A\*STAR, Singapore 138632, Singapore

\*Correspondence: [sophia.tsoka@kcl.ac.uk](mailto:sophia.tsoka@kcl.ac.uk) (S.T.); [smyiu@cs.hku.hk](mailto:smyiu@cs.hku.hk) (S.M.Y.); [fangxiang.wu@usask.ca](mailto:fangxiang.wu@usask.ca) (F.X.W.); [shaow@cnic.cn](mailto:shaow@cnic.cn) (M.X.); [s.pan@griffith.edu.au](mailto:s.pan@griffith.edu.au) (S.P.); [huiyu@nwpu.edu.cn](mailto:huiyu@nwpu.edu.cn) (H.Y.); [hes1224@163.com](mailto:hes1224@163.com) (S.H.); [xiaoli.li@sutd.edu.sg](mailto:xiaoli.li@sutd.edu.sg) (X.L.); [boxiaoc@163.com](mailto:boxiaoc@163.com) (X.B.); [wumin@a-star.edu.sg](mailto:wumin@a-star.edu.sg) (M.W.); [jianyushi@nwpu.edu.cn](mailto:jianyushi@nwpu.edu.cn) (J.Y.S.)

Received: April 29, 2025; Accepted: January 9, 2026; Published Online: January 17, 2026; <https://doi.org/10.59717/j.xinn-life.2026.100196>

© 2026 The Author(s). This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>).

Citation: Yang G., Li J., Wang Q., et al. (2026). A comprehensive survey on artificial intelligence for biomolecule design. *The Innovation Life* 4:100196.

Artificial Intelligence (AI), particularly Generative AI (GenAI), has revolutionized molecule design by enabling the creation of novel and complex molecular structures, advancing drug discovery, materials science, and synthetic biology. This survey provides a comprehensive exploration of AI-driven molecule design, categorized within and beyond the central dogma of molecular biology. We first introduce the foundational concepts of GenAI, focusing on architectures such as Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Diffusion Models (DMs), which underpin molecular design applications. Then, we examine how GenAI facilitates molecule design within the central dogma, addressing DNA, RNA, and protein design. Beyond the central dogma, we explore applications in designing small molecules, lipids, glycans, and other biomaterials. Finally, we conclude by discussing current challenges, emerging trends, and potential future directions for GenAI in molecule design. Overall, this survey aims to equip researchers with a thorough understanding of this rapidly-evolving field and inspire novel developments and breakthroughs in GenAI for molecule design.

## INTRODUCTION

The design of molecules plays a pivotal role in advancing fields such as drug discovery,<sup>1,2</sup> materials science,<sup>3</sup> and synthetic biology.<sup>4</sup> Different types of molecules, including DNAs, RNAs, proteins, small molecules, lipids, glycans, and biomaterials, serve as the foundation of these fields. For example, in drug discovery, small molecules are engineered to interact with biological targets with high specificity and precision, while proteins and RNA molecules are increasingly utilized as therapeutic agents.<sup>5–8</sup> In materials science, novel molecular structures are essential for developing functional materials with specific properties.<sup>9</sup> These diverse applications highlight the immense importance of molecule design across scientific and industrial domains.

GenAI has emerged as a powerful tool for molecule design, enabling the creation of novel and complex molecular structures.<sup>10,11</sup> AI methods can efficiently navigate vast chemical and biological spaces, automating the exploration and optimization of molecular properties. Recent advances in AI architectures, such as VAEs, GANs, and DMs, have significantly enhanced our ability to design molecules with desired functionalities and predict their interactions with biological or material systems.<sup>12,13</sup>

Molecule design strategies can be broadly classified into three categories: sequence-based design, structure-based design, and sequence-structure co-design.<sup>14–16</sup> Sequence-based design focuses on generating molecular sequences, such as DNA, RNA, or protein sequences,<sup>17–20</sup> that align with specific functional requirements. Structure-based design emphasizes gener-

ating or optimizing 3D molecular structures to achieve specific binding or stability properties.<sup>21,22</sup> Sequence-structure co-design integrates both sequence and structural information, enabling a more holistic approach to molecular engineering.<sup>23,24</sup> These strategies, while distinct, often benefit from integration, especially when combined with AI methods to enhance efficiency and precision.

Despite these advancements, AI-driven molecule design faces several challenges. Effectively exploring vast molecular spaces remains a significant scalability challenge for many AI models, leading to incomplete or suboptimal results. The accuracy of current models in predicting molecular properties and interactions remains limited, particularly for complex systems where experimental data is scarce.<sup>25,26</sup> Moreover, the interpretability of AI-generated molecules and their designs often poses a barrier to adoption in experimental workflows.<sup>27,28</sup> Future directions in this field include the development of more interpretable and efficient AI models, better integration of domain knowledge (e.g., chemical rules or biological constraints), and multi-objective optimization to balance trade-offs between molecular properties.<sup>29,30</sup> Additionally, incorporating feedback loops between AI and experimental validation can further enhance the reliability and practical impact of AI-designed molecules.

Unlike existing surveys that often focus on specific types of molecules<sup>31–38</sup> or methodologies,<sup>39–42</sup> this survey covers a broader scope, addressing both biological macromolecules such as DNAs, RNAs, and proteins, and other molecular types such as drug-like small molecules, glycans, lipids, and biomaterials. For the molecules within the central dogma, we focus on the AI-driven design and optimization of DNA regulatory elements, RNA sequences (such as RNA inverse folding and mRNA sequence optimization), and both sequences and structures for proteins. For the molecules beyond the central dogma, we discuss AI methods for generating drug-like small molecules that bind to specific targets (e.g., proteins, DNAs, or RNAs) or meet desired property criteria, as well as AI-assisted property prediction and structure design for glycans, lipids, and biomaterials. Lastly, we highlight the latest advancements in GenAI techniques and their impact on molecular design.

## GENERATIVE AI

GenAI models aim to support data synthesis by linking conceptual theories with observed data<sup>43</sup> and have been applied to various modalities, such as image,<sup>44</sup> text,<sup>45</sup> graph,<sup>46</sup> voice,<sup>47</sup> as well as sequential data,<sup>48</sup> topology data<sup>49</sup> and tabular data.<sup>50</sup> Focusing on molecule design, generative models are increasingly pivotal for the generation of novel molecules with specific functionalities and attributes through the study and simulation of complex molecular structures and properties. This section presents an overview of generative

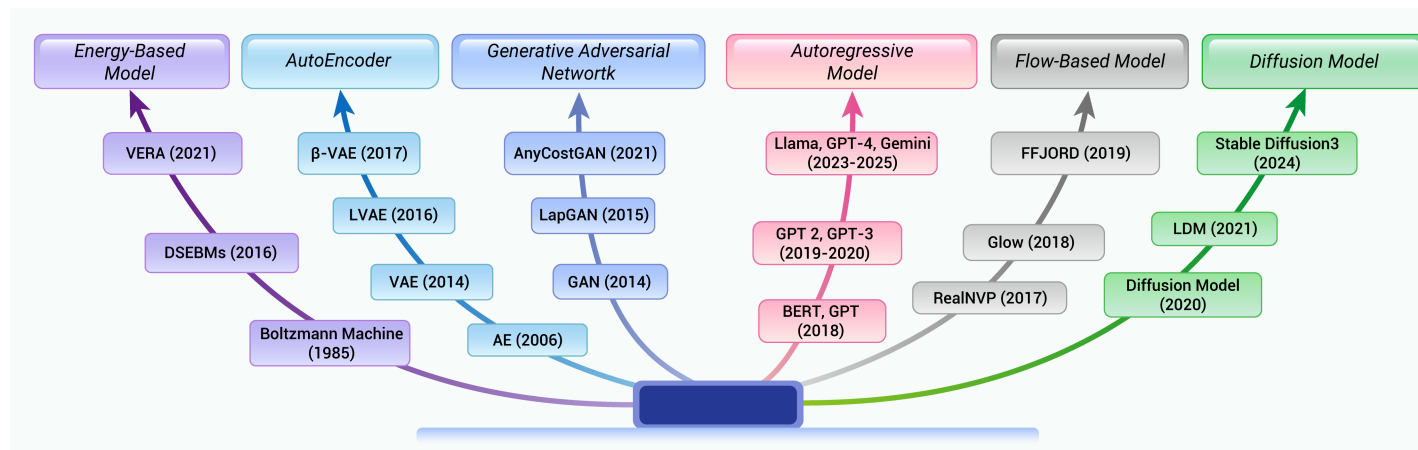


Figure 1. Landscape of GenAI Methodologies.

Table 1. Descriptions of technical terms in representative generative model frameworks.

| Model Class          | Key Term                  | Description                                                                                                                                                                                                                                 |
|----------------------|---------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Energy-Based Model   | Energy Function           | A function that assigns a scalar energy (score) to each possible configuration; lower energy corresponds to higher probability.                                                                                                             |
| VAE                  | Variational Inference     | A probabilistic framework that learns to encode data into a latent distribution (rather than fixed points) and decode back to data space, optimizing the evidence lower bound (ELBO) to balance reconstruction accuracy and regularization. |
| Flow-based Model     | Invertible Transformation | A sequence of bijective (invertible) mappings that allow exact transformation between data space and latent space, enabling exact likelihood computation.                                                                                   |
| GAN                  | Discriminator             | A neural network that tries to distinguish between real data and data generated by the generator, driving adversarial learning.                                                                                                             |
| DM                   | Diffusion Process         | A generative process that gradually adds noise to data (forward process), and then learns to reverse this process step by step to generate new samples (backward process).                                                                  |
| Autoregressive Model | Conditional Probability   | The probability of the next token or data point is modeled given its preceding context, enabling sequential generation.                                                                                                                     |

methodologies and offers a systematic review of contemporary techniques. We present a tree figure to illustrate the landscape of the included methods in Figure 1. To aid non-AI readers, we provide brief definitions of key terms in Table 1. Early energy-based models (EBMs) helped researchers understand complex distributions, leading to the idea of encoding information in lower-dimensional spaces via autoencoders (AEs). This notion naturally transitioned into adversarial training methods in GANs, further pushing the boundaries in generating realistic data. Building on these advances, DMs emerged as a novel approach to iteratively refine data from pure noise to detailed outputs. Besides those above, flow-based approaches (which utilize invertible architectures and exact likelihood computation) do not fit neatly into the evolutionary progression described above. Instead, they represent a complementary, independent branch of generative modeling that addresses challenges and offers distinct advantages, such as tractable likelihoods. Recently, autoregressive modeling has become one of the most influential paradigms in natural language processing and GenAI. Together, these approaches form the foundation of modern GenAI, each offering unique perspectives and technical solutions for biomolecule design across diverse domains.

### Energy-based approaches

EBMs have gained significant attention for their ability to model complex data distributions (e.g.,<sup>51</sup>). Boltzmann Machines (BMs) and Restricted Boltzmann Machines<sup>52</sup> (RBM) are among the earliest EBMs, employing networks of symmetrically connected visible and hidden units. Deep Belief Networks (DBNs)<sup>53</sup> stack multiple layers of RBMs to capture more complex hierarchical features. They model the probability distribution of data through an energy function optimized using techniques such as contrastive divergence, which makes them particularly suited for pattern recognition and generative modeling tasks due to their capacity to capture intricate dependencies within data. Building upon traditional BMs, Deep Structured Energy-Based Models (DSEBMs) enhance this framework by incorporating deep architectures to

represent energy functions, thereby capturing more complex, high-dimensional data distributions.<sup>54,55</sup> Similarly, the Variational Energy-Based Residual Architecture (VERA) integrates variational inference with residual structures to achieve more stable and efficient training processes. The use of residual connections in VERA helps mitigate the vanishing gradient problem often encountered in deep networks.

Typically, EBMs assign an unnormalized probability to each configuration  $x$  using an energy function:

$$p_{\theta}(x) = \frac{1}{Z_{\theta}} \exp(-E_{\theta}(x)), \quad (1)$$

where  $E_{\theta}(x)$  is the learnable energy function parameterized by  $\theta$ , and  $Z_{\theta} = \int \exp(-E_{\theta}(x)) dx$  is the partition function that normalizes the probability distribution. The training objective is typically to maximize the likelihood of real data samples  $x \sim p_{\text{data}}(x)$ , equivalently minimizing the negative log-likelihood:

$$\min_{\theta} \mathbb{E}_{x \sim p_{\text{data}}} [E_{\theta}(x)] + \log Z_{\theta}, \quad (2)$$

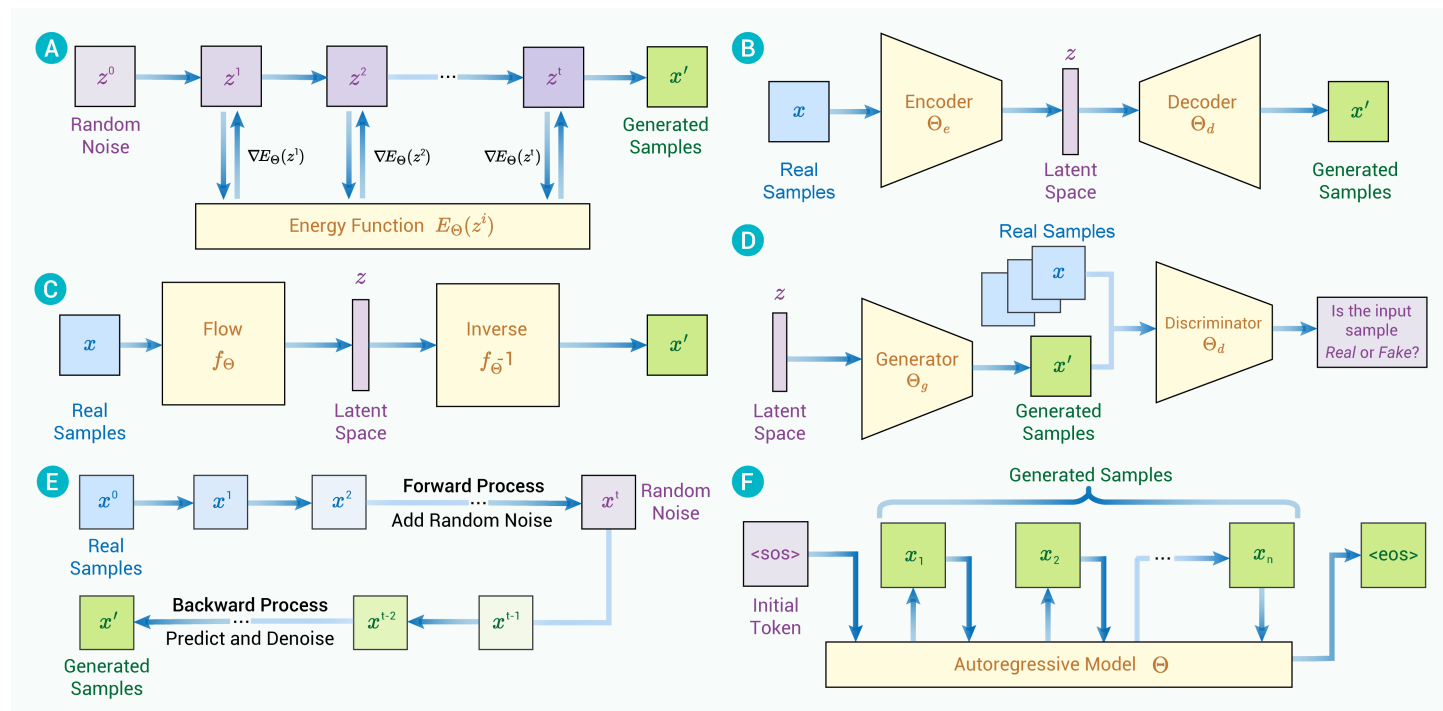
The gradient of the log-likelihood with respect to the model parameters is:

$$\nabla_{\theta} \log p_{\theta}(x) = -\nabla_{\theta} E_{\theta}(x) + \mathbb{E}_{x' \sim p_{\theta}(x')} [\nabla_{\theta} E_{\theta}(x')], \quad (3)$$

where the first term is the gradient for real data, and the second term requires sampling from the model distribution. Since direct sampling from  $p_{\theta}(x)$  is generally intractable, EBMs often employ Markov Chain Monte Carlo (MCMC) methods such as Langevin dynamics to draw approximate samples.<sup>56</sup> The iterative Langevin sampling procedure for generating a sample  $x'$  is:

$$z_{i+1} = z_i - \frac{\epsilon}{2} \nabla_z E_{\theta}(z_i) + \eta_i, \quad \eta_i \sim \mathcal{N}(0, \epsilon), \quad (4)$$

where  $z_0$  is initialized randomly (e.g., from a Gaussian),  $\epsilon$  is the step size, and



**Figure 2. Taxonomy of GenAI Models** (A) EBMs: Sequential refinement process starting from random noise  $z^0$ , iteratively updating through energy function gradients  $\nabla E_{\Theta}(z^i)$  to generate samples  $x'$ . (B) VAEs: Encoder-decoder architecture mapping real samples  $x$  to latent space  $z$  through encoder  $\Theta_e$ , then reconstructing to generated samples  $x'$  via decoder  $\Theta_d$ . (C) Flow-based Models: Bijective transformation using flow function  $f_{\Theta}$  and its inverse  $f_{\Theta}^{-1}$  to map between real samples  $x$  and latent space  $z$ , enabling exact likelihood computation. (D) GANs: Adversarial framework where generator  $\Theta_g$  produces samples  $x'$  from latent space  $z$ , competing against discriminator  $\Theta_d$  that distinguishes real from generated samples. (E) DMs: Dual-process framework with forward process progressively adding noise to real samples ( $x^0 \rightarrow x^1 \rightarrow \dots \rightarrow x^t$ ) and backward process learning to denoise and generate samples ( $x^{t-1} \rightarrow \dots \rightarrow x^0$ ). (F) Autoregressive Models: Sequential generation architecture producing tokens  $x_1, x_2, \dots, x_n$  conditioned on previous tokens, using special tokens  $\langle \text{SOS} \rangle$  and  $\langle \text{EOS} \rangle$  to mark sequence boundaries.

$x' = z^{t+1}$  after  $t$  steps. This process is illustrated in Figure 2A. In summary, the generative process of an EBM consists of sampling  $x'$  by iteratively updating latent variables using the gradient of the energy function, which guides the samples towards regions of lower energy (higher probability):

$$x' = z^{t+1}, \quad \text{where } z_{i+1} = z_i - \frac{\epsilon}{2} \nabla E_{\Theta}(z_i) + \eta_i. \quad (5)$$

This approach allows EBMs to flexibly model complex data distributions, with the learned energy landscape governing both inference and generation.

Beyond these architectural advancements, several studies have applied EBMs to *de novo* molecular generation and property optimization. For instance, GraphEBM<sup>57</sup> learns an energy function over molecular graphs and employs Langevin dynamics to sample graphs with low energy. By reducing the energy of molecules that exhibit desired properties, GraphEBM facilitates goal-directed generation. In a related approach, MolEBM<sup>58</sup> combines a VAE with a latent-space EBM prior, where the EBM jointly models molecules and their properties in a low-dimensional latent space, enabling efficient sampling of molecules that meet specified property constraints. Nevertheless, training EBMs remains a notoriously challenging problem. These models require approximating the partition function or its gradient—often via sampling methods—which can be computationally expensive and slow to mix. Additionally, ensuring training stability is difficult; if the energy landscape is poorly learned, generated samples may concentrate on a few modes or even collapse.

### Variational autoencoder approaches

AEs have played a pivotal role in dimensionality reduction and representation learning.<sup>59</sup> Originally introduced by Hinton and Salakhutdinov,<sup>60</sup> these neural network architectures are designed to learn efficient representations of input data by compressing the inputs into a latent space and subsequently reconstructing them. This process captures essential data patterns while reducing dimensionality.

An AE consists of an encoder network  $\Theta_e$  and a decoder network  $\Theta_d$ . The encoder compresses the input  $x$  into a latent representation  $z$ , while the decoder reconstructs the input from  $z$ :

$$z = \Theta_e(x), \quad x' = \Theta_d(z) \quad (6)$$

where  $x$  is a real sample,  $z$  is the latent code, and  $x'$  is the reconstructed (or generated) sample. The standard AE is trained by minimizing the reconstruction error between  $x$  and  $x'$ :

$$\min_{\Theta_e, \Theta_d} \mathbb{E}_{x \sim p_{\text{data}}} [\|x - \Theta_d(\Theta_e(x))\|^2] \quad (7)$$

Standard AEs are primarily used for dimensionality reduction and representation learning rather than generation. Their deterministic mapping ensures stable reconstructions but does not provide explicit generative capability. As shown in Figure 2B, VAEs<sup>61</sup> extend AEs with a probabilistic latent space, enabling sampling and generative applications. Specifically, the encoder outputs a distribution  $q_{\Theta_e}(z|x)$ , and the decoder defines  $p_{\Theta_d}(x|z)$ . The VAE is trained by maximizing the ELBO:

$$\max_{\Theta_e, \Theta_d} \mathbb{E}_{x \sim p_{\text{data}}} [\mathbb{E}_{z \sim q_{\Theta_e}(z|x)} [\log p_{\Theta_d}(x|z)] - \text{KL}(q_{\Theta_e}(z|x) \| p(z))] \quad (8)$$

where  $p(z)$  is a chosen prior, and the KL divergence regularizes the latent space. For generation, a latent vector  $z$  is sampled from the prior  $p(z)$ , and a new sample is generated by decoding:

$$x' = \Theta_d(z), \quad z \sim p(z) \quad (9)$$

Variational inference allows VAEs to capture complex data distributions, significantly improving generative performance.

Building upon VAEs, two representative extensions highlight different directions of improvement. Conditional Variational Autoencoders (CVAEs)<sup>62</sup> extend VAEs by incorporating conditional variables, which enables controlled or on-demand generation through side information such as class labels or molecular properties. Ladder Variational Autoencoders (LVAEs),<sup>63</sup> on the other hand, introduce a hierarchy of latent variables to capture multi-scale data features, thereby enhancing the expressiveness of latent representations and improving scalability on complex datasets. Together, these variants illustrate how VAE-based approaches can be adapted either for conditional control over the generative process or for richer hierarchical latent structure. Another influen-



tial extension is the  $\beta$ -VAE,<sup>64</sup> which introduces a tunable weight  $\beta$  on the KL-divergence term in the ELBO. By enforcing stronger regularization on the latent distribution,  $\beta$ -VAE encourages disentangled representations, making it a widely adopted framework for interpretable and controllable generative modeling. The applicability of VAE-based approaches has been broadened to various fields, including the synthesis of complex data such as molecular structures<sup>65</sup> and 3D objects.<sup>66</sup> For instance, one pioneering study<sup>67</sup> trained a VAE on molecular SMILES strings, mapping each molecule to a continuous latent vector. This continuous latent space facilitated interpolation between molecules and enabled gradient-based optimization. Moreover, Grammar VAE<sup>68</sup> incorporated chemical grammar rules directly into the decoder, ensuring that any latent point produces a syntactically valid SMILES string. This approach not only increased the fraction of valid molecules but also resulted in a smoother latent space.

Nonetheless, the quality of the latent space is crucial. Optimization paths may traverse invalid regions or encounter "dead zones"—areas that decode to non-viable structures. Integrating more intricate chemical constraints, such as specific functional groups or synthetic feasibility considerations, directly within the latent space remains challenging.

### Flow-based approaches

Flow-based generative models have emerged as a powerful class of probabilistic models for learning complex distributions, primarily due to their use of invertible transformations that enable exact likelihood computation.

As depicted in Figure 2C, a flow-based model defines an invertible mapping  $f_\theta: x \rightarrow z$  between the data space  $x \in \mathbb{R}^D$  and a latent space  $z \in \mathbb{R}^D$ . The transformation  $f_\theta$  is parameterized by neural networks with parameters  $\theta$ . The inverse mapping  $f_\theta^{-1}: z \rightarrow x$  allows for exact and efficient sampling. The change-of-variables formula enables exact computation of the data likelihood as follows:

$$\log p(x) = \log p_z(z) + \log \left| \det \frac{\partial f_\theta(x)}{\partial x} \right|, \quad \text{where } z = f_\theta(x) \quad (10)$$

Here,  $p_z$  is the prior distribution in latent space (typically a standard Gaussian), and the Jacobian determinant accounts for the volume change under the transformation. The model is trained by maximizing the log-likelihood of observed data samples  $x$ :

$$\max_{\theta} \mathbb{E}_{x \sim p_{\text{data}}} [\log p(x)] \quad (11)$$

For the generation, a latent variable  $z$  is first sampled from the prior  $p_z$ , then mapped to the data space using the inverse flow:

$$x' = f_\theta^{-1}(z), \quad z \sim p_z \quad (12)$$

This invertible structure guarantees that every generated sample  $x'$  is valid under the model, and allows for efficient and exact density estimation as well as sampling.

Among these, RealNVP<sup>69</sup> introduced a seminal approach for density estimation with invertible neural networks. By employing affine coupling layers, RealNVP facilitates efficient computation of both forward and inverse transformations, thereby allowing for exact likelihood evaluation and practical generative modeling. Building on RealNVP, Glow<sup>70</sup> integrates invertible 1×1 convolutions to mix latent dimensions more effectively while preserving the model's reversible structure. This innovation enhances expressiveness and improves the learning of complex, high-dimensional distributions. Furthermore, FFJORD<sup>71</sup> extends the flow-based framework to continuous-time dynamics through neural ordinary differential equations (ODEs). This continuous formulation offers greater flexibility for modeling distributions and capturing complex changes in the data manifold, albeit with increased computational overhead due to the requirements of numerical integration.

The success of flow-based models in text, image, and video generation highlights their ability to capture complex, high-dimensional distributions. Their potential for molecule design arises from structural analogies between tasks. For example, generating molecular sequences parallels natural language generation, where both involve modeling discrete sequential data. Similarly, generating molecular structures corresponds to graph generation,

as molecules can be represented as nodes and edges in a graph. For instance, GraphNVP<sup>72</sup> adapts the flow-based approach to molecular graphs by learning a continuous latent space for molecular structures, enabling the generation of diverse and chemically valid small molecules. Additionally, extending these methods to larger biomolecular systems, the Proteina model<sup>73</sup> employs conditional generation techniques to design protein backbones. By conditioning on structural features such as fold type, Proteina facilitates controlled *de novo* protein design.

In summary, flow-based models provide an elegant probabilistic framework with exact likelihood evaluation and efficient sampling. Their extensions to molecular graphs and protein sequences highlight their versatility for molecular design, although challenges such as scalability and domain-specific constraints remain active research directions.

### Generative adversarial network

GANs<sup>74</sup> have revolutionized image synthesis through an adversarial framework that enables the creation of highly realistic images.

Figure 2D provides an illustration of the original GAN framework,<sup>75</sup> which could be formulated as a two-player minimax game between a generator  $\theta_g$  and a discriminator  $\theta_d$ , where the generator learns to map samples  $z$  from a prior distribution  $p_z$  (e.g., Gaussian noise) to the data space, and the discriminator aims to distinguish between real data samples  $x \sim p_{\text{data}}$  and generated samples  $x' = \theta_g(z)$ . Formally, the objective function is given as:

$$\min_{\theta_g} \max_{\theta_d} V(\theta_g, \theta_d) = \mathbb{E}_{x \sim p_{\text{data}}} [\log \theta_d(x)] + \mathbb{E}_{z \sim p_z} [\log(1 - \theta_d(\theta_g(z)))] \quad (13)$$

Here,  $\theta_d(x)$  represents the probability that  $x$  is a real sample rather than a fake one. The generator  $\theta_g(z)$  seeks to produce samples that can "fool" the discriminator, while the discriminator  $\theta_d$  tries to maximize its accuracy. During training,  $\theta_g$  and  $\theta_d$  are updated alternately: the discriminator is trained to maximize the likelihood of correctly identifying real and generated samples, while the generator is trained to minimize  $\log(1 - \theta_d(\theta_g(z)))$ , or equivalently, to maximize  $\log \theta_d(\theta_g(z))$  for better gradients.<sup>75</sup> Once training converges, generation is performed by sampling a latent variable  $z \sim p_z$  and passing it through the generator to obtain a new sample:  $x' = \theta_g(z)$ . In molecular generation,  $\theta_g$  can be designed to output molecular graphs, SMILES strings, or protein sequences, depending on the task. Building upon the original framework, Conditional GANs (CGANs)<sup>76</sup> marked a pivotal milestone by conditioning the generation process on auxiliary information such as class labels. This innovation transformed GANs from purely random generators into controllable generative models, thereby enabling on-demand sample generation in both vision and molecule design. Beyond CGANs, numerous specialized variants have emerged. For instance, LapGAN<sup>77</sup> employs a Laplacian pyramid structure to enhance image super-resolution, refining images in a coarse-to-fine manner by progressively predicting high-frequency details. AnyCostGAN<sup>78</sup> addresses the high computational costs of large-scale GANs (e.g., StyleGAN2<sup>79</sup>) by offering flexible computation with variable resolutions and channel configurations, thus balancing efficiency with output quality. Rather than a linear evolution, these developments illustrate diverse adaptation directions: CGANs emphasize controllability, LapGANs improve high-resolution synthesis, and AnyCostGANs enhance computational scalability.

These advancements have also been extended to molecular generation, where molecular graphs or sequences are treated analogously to images. For instance, MolGAN<sup>80</sup> generates molecular graphs—comprising node and edge features—using a generator network, with a discriminator network distinguishing generated graphs from real ones. In the domain of larger biomolecules, GANs have been applied to protein design. ProteinGAN<sup>81</sup> is a notable example that trains on natural protein sequences; its generator produces amino acid sequences, while the discriminator guides the model to generate sequences that are indistinguishable from genuine proteins.

Despite these sophisticated techniques, GAN training remains notoriously unstable—a challenge that persists in molecular applications. However, with innovations such as improved objective functions (e.g., WGAN and energy-based GANs) and enhanced graph representations, GANs continue to be explored for tasks requiring high-realism sample generation, particularly where valid syntax is assured and the primary challenge is to capture functional diversity.

### Diffusion approaches

DMs<sup>82</sup> have emerged as powerful generative frameworks characterized by a stepwise noise reduction process that transforms random noise into coherent data samples. As depicted in Figure 2E, classic DMs<sup>83,84</sup> define a forward process that gradually adds Gaussian noise to the data, which can be defined as:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_0, (1 - \alpha_t)\mathbf{I}), \quad (14)$$

where  $x_0$  is the original real sample,  $x_t$  is the noisy sample at step  $t$ , and  $\alpha_t$  is a noise schedule. More generally, the diffusion process can be described by a stochastic differential equation (SDE):

$$dx = f(x, t)dt + g(t)dw, \quad (15)$$

where  $f(x, t)$  and  $g(t)$  control the drift and diffusion, and  $dw$  denotes the Wiener process<sup>85</sup> (i.e., standard Brownian motion). The reverse process, used for generation, is learned by estimating the score function  $\nabla_{x_t} \log q(x_t)$ . For instance, score matching<sup>84</sup> minimizes the difference between the true and estimated gradients of the data distribution, typically through the following objective:  $\mathbb{E}_{x_0, t, x_t} [\lambda(t) |\nabla_{x_t} \log q(x_t|x_0) - s_\theta(x_t, t)|^2]$ , where  $s_\theta$  is a neural network to parameterize the score function.

While early approaches enable flexible and high-quality sample generation, they typically require significant computational resources due to the many steps involved. Latent Diffusion Models (LDMs)<sup>86</sup> mitigate this computational burden by performing the diffusion process in a compressed latent space rather than in pixel space. This approach reduces computational load and accelerates the generation process while maintaining high output quality. More recently, Stable Diffusion 3<sup>87</sup> has further refined the latent diffusion framework by optimizing model architecture and training methods to enhance speed and scalability, thereby enabling real-time image synthesis and editing. For molecular generation, DMs have also shown promise in generating both small and large molecules. An early example in chemistry is GraphDiffusion,<sup>88</sup> which introduced a score-based generative model for molecular graphs. GraphDiffusion demonstrated that DMs can generate molecular graphs with quality on par with or even superior to that of GANs and VAEs on common datasets, marking the first use of continuous diffusion for graph generation. Subsequent advances include the E(3)-equivariant diffusion model (EDM),<sup>89</sup> which diffuses both atom coordinates (continuous 3D positions) and atom types. EDM employs a Euclidean symmetry-equivariant network to learn the reverse process, enabling the high-quality generation of molecules with accurate 3D geometries and achieving state-of-the-art results in producing stable molecular conformations.

In summary, DMs tend to generate high-quality, diverse samples by explicitly learning the underlying data distribution via denoising score matching. They are less prone to mode collapse than GANs and can better cover the breadth of molecular space. However, a notable drawback is their computational intensity, as generating a single molecule typically requires tens or even hundreds of denoising steps, making them considerably slower than one-shot generation methods.

### Autoregressive approaches

Autoregressive models<sup>90,91</sup> are a cornerstone of modern generative artificial intelligence, forming the theoretical backbone for recent breakthroughs such as Large Language Models (LLMs) and other foundation models. The foundation for most contemporary autoregressive models is the Transformer architecture.<sup>92</sup> Its self-attention mechanism allows the model to weigh the importance of different tokens in the input sequence, enabling the capture of long-range dependencies far more effectively than previous recurrent architectures.

The introduction of the Transformer<sup>92</sup> paved the way for a series of powerful autoregressive models. The Generative Pre-Training (GPT) family, starting with GPT-3<sup>93</sup> established the paradigm of large-scale unsupervised pre-training for language understanding. While its contemporary, bidirectional encoder representation from transformers (BERT),<sup>94</sup> also belongs to the Transformer family, it is a bidirectional auto-encoding model designed for understanding tasks, not sequential generation. The subsequent releases of GPT-2<sup>45</sup> and especially GPT-3,<sup>95</sup> with its unprecedented scale of over 175 billion parameters,

demonstrated emergent "few-shot" and "zero-shot" learning capabilities. This lineage continued to evolve with GPT-4<sup>96</sup> and GPT-4o,<sup>97</sup> which integrated strong multimodal capabilities for processing and generating text, images, and audio. Other significant models like XLNet<sup>98</sup> sought to combine the strengths of autoregressive (GPT-like) and auto-encoding (BERT-like) models, while GLM<sup>99</sup> introduced an innovative "autoregressive blank infilling" objective. The open-source community has also been pivotal, with releases like Meta's OPT<sup>100</sup> and LLaMA series,<sup>101,102</sup> which have democratized access to powerful large models.

Figure 2F depicts the generation process of the autoregressive model. Mathematically, an autoregressive model parameterized by  $\theta$  models the joint probability distribution of a sequence of tokens  $x = (x^1, x^2, \dots, x^T)$  by factorizing it into a product of conditional probabilities:

$$P(x; \theta) = \prod_{t=1}^T P(x^t | x^1, \dots, x^{t-1}; \theta) \quad (16)$$

The model is trained by maximizing the log-likelihood of a training dataset, which corresponds to minimizing the negative log-likelihood loss function  $L(\theta)$ :

$$L(\theta) = -\sum_{i=1}^N \log P(x_i; \theta) = -\sum_{i=1}^N \sum_{t=1}^{T_i} \log P(x_i^t | x_i^{<t}; \theta) \quad (17)$$

During inference, a new sequence is generated step-by-step. Starting with an initial prompt (or a start-of-sequence token), the model samples the next token  $x^t$  from the conditional probability distribution  $P(\cdot | x^1, \dots, x^{t-1}; \theta)$ . This newly generated token is then appended to the sequence and fed back into the model to generate the subsequent token, repeating until an end-of-sequence token is produced or a predefined length is reached.

From an AI for biomolecule design perspective, autoregressive models have been successfully applied to generate chemical structures, protein sequences, and reaction pathways by treating molecular representations (such as SMILES strings) as token sequences.<sup>25</sup> This enables the generation of novel molecules with desired properties, driven by the expressive power of large-scale autoregressive Transformers. Despite their success, autoregressive models have notable limitations. They typically require massive amounts of data and computational resources for effective training, which can be a bottleneck in domains with limited annotated data. Additionally, the sequential generation process can be slow, and the models may struggle with long-term dependencies beyond their context window.

### Summary

The diverse GenAI technologies discussed above—EBMs, AEs, flow-based models, GANs, DM, and autoregressive models—each offer distinct capabilities for molecular design, with complementary strengths and limitations that make them suitable for different applications. Table 2 provides a systematic comparison across four key performance dimensions. DMs achieve the highest generation quality and diversity, establishing them as the current state-of-the-art despite computational speed limitations.<sup>89</sup> VAEs demonstrate exceptional speed and stability with fair generation quality, making them highly practical for molecular property optimization.<sup>67</sup> Flow-based models excel in training stability while maintaining good quality and diversity through invertible transformations,<sup>72</sup> but face speed constraints due to architectural limitations. GANs achieve good generation quality with excellent speed, enabling rapid high-fidelity sampling,<sup>80</sup> yet suffer from training instability that limits practical deployment. EBM models provide good quality and diversity with theoretical flexibility for complex distribution modeling, but computational bottlenecks due to MCMC sampling limit their practical application.<sup>56</sup> Autoregressive models offer balanced performance across all dimensions, providing reliable and controllable generation leveraging advances in LLMs.<sup>25</sup>

To better understand the practical implementation and evolution of GenAI technologies in molecular design, we categorize their application progress into six progressive stages reflecting AI's increasing sophistication. Stage 1 (Prediction-Assisted) uses AI models as predictive tools for human-driven design through property prediction. Stage 2 (Optimization-Guided) leverages AI optimization to guide modification of existing structures. Stage 3 (Partial Generative) introduces generative capabilities for creating molecular fragments.

Table 2. Comparative analysis of GenAI approaches for biomolecule design. Rating scale: - (Poor), + (Fair), ++ (Good), +++ (Excellent).

| Model          | Quality | Speed | Stability | Diversity | Key Strength             | Main Limitation          |
|----------------|---------|-------|-----------|-----------|--------------------------|--------------------------|
| Energy-Based   | ++      | -     | +         | ++        | Flexible distributions   | Slow MCMC sampling       |
| VAE            | +       | +++   | +++       | +         | Stable & interpretable   | Latent space quality     |
| Flow-Based     | ++      | +     | +++       | ++        | Exact likelihood         | Architecture constraints |
| GAN            | ++      | +++   | -         | +         | High-fidelity samples    | Training instability     |
| Diffusion      | +++     | -     | ++        | +++       | State-of-the-art quality | Multi-step sampling      |
| Autoregressive | ++      | +     | ++        | ++        | Sequential control       | Element-wise generation  |

Stage 4 (Unconditional Generative) achieves complete molecule generation with autonomous capacity but lacks targeted control. Stage 5 (Conditional Generative) enables constraint-based generation for targeted design according to specific requirements. Stage 6 (Autonomous Design) represents the ultimate goal of intelligent, interpretable creation and optimization under complex multi-objective constraints. Table 3 meticulously outlines the progress of key molecular design tasks—including DNA, RNA, protein, small molecules, glycans and lipids, and biomaterials—across different AI application stages (from prediction-assisted to conditional generative, etc.).

This staged framework not only reflects the progressive enhancement of AI technological capabilities but also reveals the evolutionary transition from auxiliary tools to intelligent design agents. Ensuring chemical validity, synthetic feasibility, and interpretability of generated molecules remains a critical hurdle, while multi-objective optimization of properties like biological activity and toxicity adds further complexity. These challenges highlight the ongoing need for innovation to enhance the utility, reliability, and practical application of GenAI in molecular design. In the following sections, we will explore how these stages manifest across different biomolecular domains—DNA, RNA, proteins, and other molecules—and discuss the substantial challenges they face within each application context.

### DNA DESIGN

DNA design focuses on the design and optimization of DNA regulatory elements, which are specific DNA sequences with regulatory functions and play a crucial role in gene expression regulation and cellular differentiation processes. Promoters initiate transcription with tunable strength and specificity. Enhancers amplify gene expression through long-range chromatin interactions. Additional elements include terminators that control transcription cessation, silencers that suppress expression, and insulators that establish genomic boundaries. Effective design requires integrating sequence pattern recognition with mechanistic understanding of protein-DNA interactions, chromatin structure, and regulatory networks. Recent advances in GenAI have transformed regulatory DNA design, enabling the creation of functional sequences with desired properties through data-driven learning from natural regulatory elements.<sup>103–105</sup>

As shown in Figure 3A, cis-acting DNA elements provide sequence-level control of transcription, while trans-acting factors, which include transcription factors (TFs) and other regulatory proteins, regulate through protein-DNA interactions. This dual perspective highlights that AI-driven DNA design must consider both intrinsic sequence features and extrinsic regulatory proteins to achieve precise control of gene expression.

### Promoter design

A promoter is a crucial DNA region that initiates gene transcription, with its sequence activity directly dictating the level of gene expression. Consequently, designing synthetic promoters with desired strengths is a cornerstone of synthetic biology. While early computational efforts focused on predicting the activity of existing sequences, the advent of GenAI has shifted the paradigm towards the *de novo* design of entirely new promoter sequences. This evolution has been marked by the application of increasingly sophisticated generative models, each with distinct characteristics and capabilities.<sup>106–109</sup>

GANs were among the first deep learning models applied to this challenge. While standard GANs can suffer from training instability and mode collapse,

variants like Wasserstein GANs (WGANs) have been developed to address these limitations. In a seminal work, Wang et al.<sup>106</sup> proposed a framework using a WGAN to learn the sequence features of natural *E. coli* promoters and generate novel, functional sequences. Building on this foundation, subsequent research like DeepSEED<sup>110</sup> has leveraged CGANs for more controllable design, enabling users to specify core elements (e.g., transcription factor binding sites) while generating compatible flanking sequences through expert-guided generative modeling. VAEs represent another class of generative models for this task, with frameworks like that of Seo et al.<sup>111</sup> combining VAE-based sequence generation with convolutional neural network (CNN)-based strength prediction to establish the effective 'generator-predictor' pipeline.

More recently, DMs have emerged as a powerful alternative due to their training stability and high-quality sample generation. These models learn to reverse a gradual noising process, iteratively refining random noise into coherent sequences. Representative works include PromoterDiff,<sup>107</sup> Lin et al.'s diffusion-based framework,<sup>112</sup> and D3,<sup>103</sup> a score-entropy discrete diffusion model operating directly in nucleotide space that produces sequences nearly indistinguishable from natural DNA. Complementing this trend, transfer learning approaches like PromoGen<sup>113</sup> leverage pre-trained language models, enabling functional promoter generation for understudied organisms through multi-species pre-training followed by species-specific fine-tuning.

Modern approaches increasingly employ hybrid systems that integrate generative and predictive components. In these architectures, generative models explore sequence space while predictive models (typically CNNs) serve as *in silico* filters, evaluating candidates for desired properties.<sup>114</sup> This synergy achieves efficient exploration of sequence space by prioritizing sequences that balance novelty with functional potential. Platforms like GPro<sup>108</sup> standardize these workflows, making advanced design capabilities accessible to the broader research community.

### Enhancer design

An enhancer is a regulatory DNA sequence that can boost gene transcription from a distance by binding TFs and forming chromatin loops to contact target promoters, regardless of its position or orientation. These regulatory elements play crucial roles in cell-type-specific and developmental stage-specific gene activation.<sup>273</sup> The *de novo* design of enhancers with desired activity and specificity is therefore a central goal in synthetic biology and gene therapy. Initial deep learning applications in this domain leveraged predictive models as *in silico* oracles to guide sequence design. For instance, DeepSTARR employed a CNN to screen billions of random sequences, identifying synthetic enhancers matching the activity of strong native variants.<sup>115</sup> While groundbreaking, this random screening approach explores the vast design space inefficiently. Subsequent works pursued more directed strategies. DREAM<sup>116</sup> coupled predictive modeling with genetic algorithms, while others employed deep learning-guided *in silico* evolution for more effective sequence space navigation.<sup>274</sup> A key advancement in this oracle-guided paradigm is the iterative design cycle, where models are retrained on experimentally validated synthetic data, leading to progressively superior designs with enhanced cell-type specificity, as demonstrated by Yin et al.<sup>117</sup>

In order to address the limitations of local optima in iterative optimization, recent approaches employ generative models that learn the underlying distribution of functional enhancers directly. These models generate novel, diverse, and biologically plausible sequences from scratch, offering a fundamentally different design approach. Recent architectures span long-context autore-

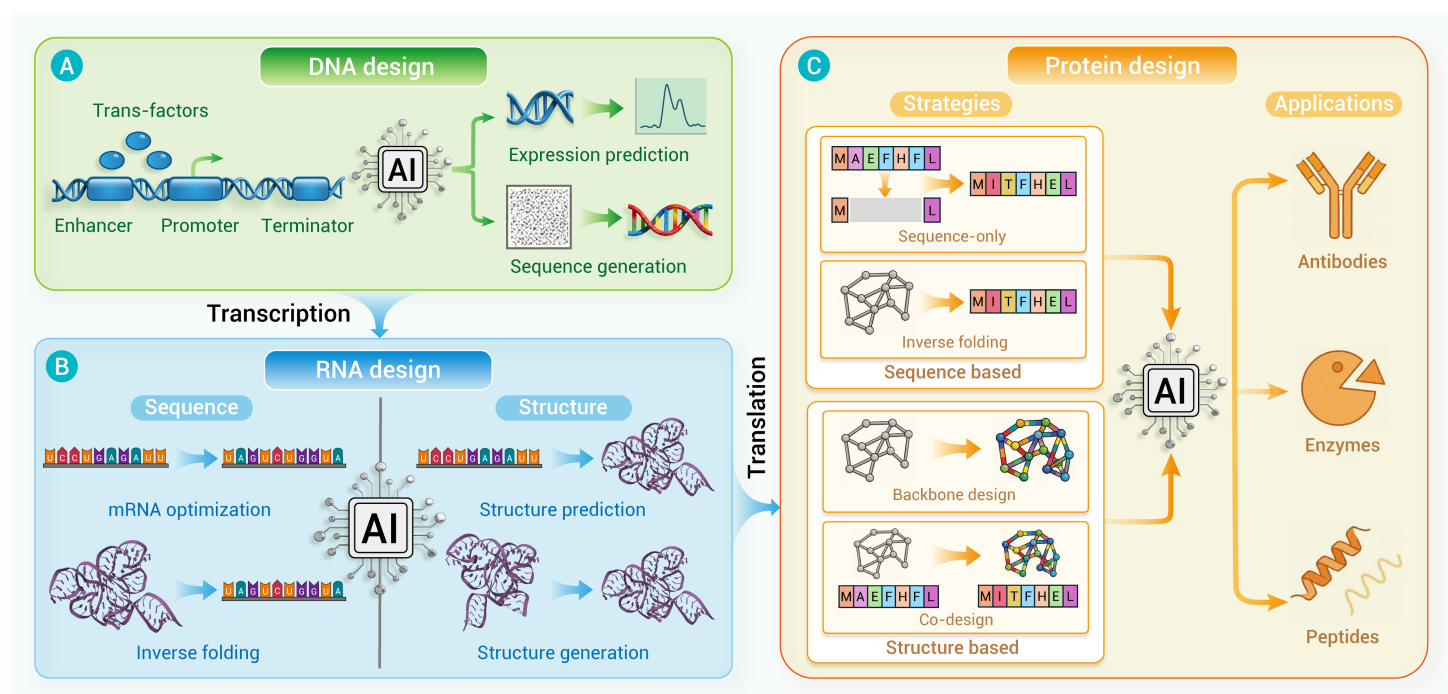
Table 3. Summary of AI applications in key biomolecule design tasks.

| Category                 | Design Task                | AI Application Stage | Representative Models                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | Key Contribution & Evaluation                                                                                                                                                        |
|--------------------------|----------------------------|----------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| DNA                      | Promoter                   | Stage 5              | Wang et al., <sup>106</sup> DeepSEED, <sup>110</sup> Seo et al., <sup>111</sup> PromoterDiff, <sup>107</sup> Lin et al., <sup>112</sup> D3, <sup>103</sup> PromoGen, <sup>113</sup> Wang et al., <sup>114</sup> GPro. <sup>108</sup>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | Achieved <i>de novo</i> design of functional promoter sequences with statistical characteristics highly similar to natural promoters.                                                |
|                          | Enhancer                   | Stage 5              | DeepSTARR, <sup>115</sup> DREAM, <sup>116</sup> Yin et al., <sup>117</sup> GENERator, <sup>104</sup> D3, <sup>103</sup> TACO, <sup>118</sup> DRAKES. <sup>119</sup>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | Achieved <i>de novo</i> design and optimization of enhancer sequences with cell-type specificity, with activity comparable to or exceeding natural variants.                         |
|                          | Terminator                 | Stage 3              | iTerm-PseKNC, <sup>120</sup> DISA, <sup>121</sup> Li et al.. <sup>122</sup>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              | Advanced AI application from activity prediction to generative design, with high termination efficiency of generated sequences verified by experiments.                              |
|                          | Silencer                   | Stage 1              | SilencerDB, <sup>123</sup> SilenceREIN. <sup>124</sup>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | Provides key information for subsequent functional validation and assisted design by accurately identifying candidate silencer elements from complex data.                           |
|                          | Insulator                  | Stage 1              | No mature representative design models yet.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              | Primarily established the functional dependency on 3D chromatin structure, pointing the direction for future assisted design.                                                        |
|                          | TFs                        | Stage 1              | DeepTFactor, <sup>125</sup> Liu et al., <sup>126</sup> CellCall, <sup>127</sup> Alerasool et al.. <sup>128</sup> ,                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | Primarily established the functional dependency on 3D chromatin structure, pointing the direction for future assisted design.                                                        |
| RNA                      | RNA inverse folding        | Stage 5              | gRNAde, <sup>129</sup> RhoDesign, <sup>130</sup> Li et al., <sup>131</sup> Runge, <sup>132</sup> RiboDiffusion. <sup>133</sup>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | Achieved <i>de novo</i> sequence design for functional RNAs (e.g., ribozymes, riboswitches), enabling them to fold into specific functional structures.                              |
|                          | mRNA sequence optimization | Stage 5              | GEMORNA, <sup>134</sup> GenerRNA, <sup>135</sup> RNAtTranslator, <sup>136</sup> RNA-BANg, <sup>137</sup> EvoFlow-RNA, <sup>138</sup> RNACG, <sup>139</sup> Castillo et al.. <sup>140</sup>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | Achieved comprehensive optimization and generation of mRNA sequences (covering UTRs and coding regions) to enhance their stability and protein expression efficiency.                |
|                          | RNA structures generation  | Stage 5              | RNAFlow. <sup>141</sup>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | Advanced AI application from predicting folded structures of existing RNAs to the <i>de novo</i> design of novel 3D RNA structures with desired functions.                           |
| Protein                  | Antibody design            | Stage 5              | ESM-1v, <sup>142</sup> RFdiffusion, <sup>143</sup> Parkinson et al., <sup>144</sup> Mason et al., <sup>145</sup> Makowski et al., <sup>146</sup> Shanehsazzadeh et al., <sup>147</sup> Shanker et al., <sup>148</sup> GUIDE, <sup>149</sup> PALM-H3, <sup>150</sup> Ig-VAE. <sup>151</sup>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | Achieved affinity optimization and <i>de novo</i> design of antibodies against specific tumor or pathogen epitopes.                                                                  |
|                          | Peptide design             | Stage 5              | AMPlify, <sup>152</sup> AMP-BERT, <sup>153</sup> Losin et al., <sup>154</sup> Sharma et al., <sup>155-156</sup> Teimouri et al., <sup>157</sup> Vishnepolsky et al., <sup>158</sup> Sharma et al., <sup>155</sup> Losin et al., <sup>154</sup> Salem et al., <sup>159</sup> Ansari et al., <sup>160-161</sup> Salem et al., <sup>159</sup> Ferrell et al., <sup>162</sup> Van et al., <sup>163</sup> Buehler et al., <sup>164</sup> HydrAMP, <sup>165</sup> CLaSS, <sup>166</sup> GFlowNets, <sup>167</sup> Gin et al., <sup>168</sup> Szymczak et al., <sup>169</sup> Hoffman et al., <sup>170</sup> Liu et al., <sup>171</sup> AOPAnOxPePred, <sup>172</sup> AnOxPP, <sup>173</sup> Hesamzadeh et al., <sup>174</sup> AntiCP, <sup>175</sup> ACP-DL, <sup>176</sup> ACPred-LAF, <sup>177</sup> xDeep-AcPEP, <sup>178</sup> xACP-MLC, <sup>179</sup> ACPScanner, <sup>180</sup> DUO-ACP, <sup>181</sup> Grisoni, <sup>182</sup> GANDALF. <sup>183</sup> | Achieved the prediction, optimization, and generation of peptide sequences with specific biological activities (e.g., antimicrobial, anticancer).                                    |
|                          | Enzyme design              | Stage 5              | Dou, <sup>184</sup> Russ et al., <sup>185</sup> Yeh, <sup>186</sup> TurboPETase, <sup>187</sup> Cui et al., <sup>188</sup> crawshaw et al., <sup>189</sup> ProteinGAN, <sup>81</sup> MutCompute, <sup>190</sup> Lauko et al.. <sup>191</sup>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | Achieved significant enhancement of the catalytic efficiency of existing enzymes and successfully designed <i>de novo</i> enzyme molecules with novel folds and catalytic functions. |
| Drug-like small molecule | Protein-guided             | Stage 5              | Grechishnikova et al., <sup>192</sup> Alphadrug, <sup>193</sup> DeepTarget, <sup>194</sup> CDGCN, <sup>195</sup> PCMoI, <sup>196</sup> LiGAN, <sup>197</sup> DESERT, <sup>198</sup> EGCM, <sup>199</sup> GAT-VAE, <sup>200</sup> CProMG, <sup>201</sup> Flag, <sup>202</sup> DrugGPS, <sup>203</sup> Targetdiff, <sup>204</sup> DecompDiff, <sup>205</sup> DeepICL, <sup>16</sup> NCIDiff. <sup>206</sup>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | Achieved <i>de novo</i> design of molecules for specific protein binding pockets, successfully advancing an AI-designed drug candidate to clinical trials.                           |
|                          | Nucleic acid-guided        | Stage 1              | RNAPosers, <sup>207</sup> AnnapuRNA, <sup>208</sup> Haga et al., <sup>209</sup> Szulc et al., <sup>210</sup> Krishnan et al., <sup>211</sup> Grimberg et al.. <sup>212</sup>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | Provides informational support for rational drug design by predicting small molecule binding modes and affinities to nucleic acid targets.                                           |
|                          | Ligand-guided molecule     | Stage 5              | Segler et al., <sup>213</sup> TSMMG, <sup>214</sup> 3DSMILES-GPT, <sup>215</sup> Sc2Mol, <sup>216</sup> TGM-DLM, <sup>217</sup> JT-VAE, <sup>218</sup> MolGAN, <sup>80</sup> GraphAF, <sup>219</sup> GraphXForm, <sup>220</sup> DiGress, <sup>221</sup> HGLDM, <sup>222</sup> SyCo, <sup>223</sup> EDM, <sup>89</sup> GeoLDM, <sup>224</sup> MolDiff, <sup>225</sup> MD3MD, <sup>226</sup> MLPS, <sup>227</sup> MCMG, <sup>228</sup> TANGO, <sup>229</sup> SSSVAE, <sup>230</sup> .                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | Achieved the exploration of vast chemical space to generate novel molecules with chemical validity and specific optimization objectives.                                             |
|                          | Molecular modification     | Stage 3              | DeLinker, <sup>231</sup> SyntaLinker, <sup>232</sup> SyntaLinker-Hybrid, <sup>233</sup> 3DLinker, <sup>234</sup> Graph-GMVAE, <sup>235</sup> DeepHop, <sup>236</sup> DiffHop, <sup>237</sup> DeepFrag, <sup>238</sup> DEVELOP, <sup>239</sup> STRIFE, <sup>240</sup> D3FG, <sup>241</sup> JT-VAE, <sup>218</sup> MoFlow, <sup>242</sup> GCPN, <sup>243</sup> GA-DNN. <sup>244</sup>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | Achieved fine-grained modification of existing lead compounds (e.g., scaffold hopping, fragment replacement) to optimize their druggability.                                         |



Table 3. (Continued)

| Category         | Design Task             | AI Application Stage | Representative Models                                                                                                                                                                                                                                                                                                                                                       | Key Contribution & Evaluation                                                                                                                                                                            |
|------------------|-------------------------|----------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Glycans & Lipids | Glycans                 | Stage 1              | GlyNet, <sup>245</sup> SweetTalk, <sup>246</sup> glyBERT, <sup>247</sup> SweetNet, <sup>248</sup> DeepGlyco, <sup>249</sup> CandyCrunch. <sup>250</sup>                                                                                                                                                                                                                     | Provides a basis for functional validation and subsequent assisted design by predicting glycan biological activities and elucidating their complex structures.                                           |
|                  | Lipids                  | Stage 5              | Han et al., <sup>251</sup> Mak et al., <sup>252</sup> LipidCCS, <sup>253</sup> adapted Synthesis-DAGs. <sup>254</sup>                                                                                                                                                                                                                                                       | Advanced AI application from property prediction to the <i>de novo</i> generative design of synthesizable lipids, achieving iterative optimization for specific functions like mRNA delivery efficiency. |
| Biomaterials     | Hydrogels               | Stage 5              | Soleimani et al., <sup>255</sup> Karaoglu et al., <sup>256</sup> Wang et al., <sup>257-258</sup> Khalvandi et al., <sup>259</sup> Zheng et al., <sup>260</sup> Islamkulov et al., <sup>261</sup> Boztepe et al., <sup>262</sup> AI-HTPCSS, <sup>263</sup> Liu et al., <sup>264</sup> Umar et al., <sup>265</sup> Qiu et al., <sup>266</sup> Cadamuro et al., <sup>267</sup> | Achieved a full loop from predicting key properties (e.g., swelling, mechanical behavior) to the inverse design of functional structures.                                                                |
|                  | Other Polymers & Metals | Stage 1              | Epa et al., <sup>268</sup> Kwaria et al., <sup>269</sup> Rostam et al., <sup>270</sup> Wu et al., <sup>271</sup> Chang et al., <sup>272</sup>                                                                                                                                                                                                                               | Provides guidance for material screening and optimization design by predicting the biological effects of polymers and key properties of metal alloys.                                                    |



**Figure 3. Design of biomolecules within the central dogma** (A) DNA design encompasses regulatory elements (enhancers, promoters, terminators) and trans-acting factors, enabling expression prediction and sequence generation. (B) RNA design spans sequence-level optimization (mRNA optimization, inverse folding) and structure-level tasks (structure prediction and generation). (C) Protein design strategies range from sequence-based (sequence-only design, inverse folding) to structure-based approaches (backbone design, sequence-structure co-design), with applications in antibodies, enzymes, and peptides.

gressive transformers, such as GENERator,<sup>104</sup> discrete DMs like D3, which operate directly in nucleotide space,<sup>103</sup> and simulation-free frameworks, including Dirichlet Flow Matching.<sup>275</sup> These approaches excel at capturing complex sequence grammar, enabling comprehensive exploration of the functional enhancer space.

Generative models can be fine-tuned with reinforcement learning (RL) to optimize enhancer design, balancing biological realism with targeted performance goals. TACO exemplifies this strategy by applying RL to enhance autoregressive models for enhancer design,<sup>118</sup> while DRAKES optimizes discrete DMs for similar enhancer generation tasks.<sup>119</sup> These enhancer design methods demonstrate how GenAI can both replicate and improve upon natural regulatory elements, enabling controlled therapeutic applications.

### Other DNA regulatory elements

Beyond promoters and enhancers, computational approaches are extending to diverse regulatory elements with varying degrees of AI integration. Termini-

nators represent the most advanced area, progressing from prediction models like iTerm-PseKNC<sup>120</sup> and design frameworks such as DISA<sup>121</sup> to true generative approaches. Li et al.<sup>122</sup> demonstrated comprehensive *de novo* design by combining strength prediction ( $R^2 = 0.72$ ) with WGAN-GP generation, achieving 72% experimental success with >90% termination efficiency.

Silencers and insulators remain primarily in prediction phases despite their regulatory importance. Silencer identification has evolved from chromatin-based approaches<sup>276</sup> to machine learning methods<sup>277</sup> and deep learning models, such as SilencerDB,<sup>123</sup> as well as multi-modal approaches like SilenceREIN.<sup>124</sup> Insulators present greater challenges due to context-dependent CTCF-mediated functions<sup>278</sup> and three-dimensional chromatin organization requirements.<sup>105</sup>

TFs occupy a distinct computational space as protein regulators rather than DNA sequences. Current AI applications focus on functional prediction through tools like DeepTFactor,<sup>125</sup> clinical applications,<sup>126</sup> and systems-level analysis via approaches like CellCall.<sup>127</sup> Structural methods leveraging

AlphaFold2 predictions enable the discovery of novel interactions,<sup>128</sup> although the rational design of synthetic TFs remains limited, representing a significant frontier for future AI-driven protein design approaches.

## Summary

GenAI has initiated a technological revolution in DNA regulatory element design, achieving breakthrough progress in promoter and enhancer generation while revealing significant opportunities for broader applications. Current successes demonstrate AI's capability to create functional regulatory sequences with desired properties through learning from massive genomic datasets. However, substantial challenges persist for other regulatory elements, including terminators, silencers, and insulators, primarily due to limited high-quality training data and incomplete understanding of their context-dependent mechanisms.<sup>277</sup> These elements often rely on complex three-dimensional chromatin interactions,<sup>115</sup> making their regulatory patterns difficult to capture with existing algorithms. Additionally, practical constraints, including synthesis feasibility, sequence compatibility, and experimental validation bottlenecks, currently limit the translation from computational design to functional implementation.

The future potential of GenAI in DNA regulation extends far beyond current capabilities. Multi-element collaborative design promises tissue-specific gene switches for precision medicine, environmentally responsive regulatory networks for adaptive cellular systems, and cross-species universal elements for biotechnology applications. Most transformatively, GenAI may enable "fully artificial genome" design through coordinated optimization of thousands of regulatory elements, creating synthetic biological systems with unprecedented functionality. Realizing these applications requires addressing three critical challenges: establishing comprehensive multi-omics databases spanning diverse regulatory elements, developing next-generation algorithms incorporating biophysical constraints and chromatin dynamics, and creating high-throughput synthesis-validation platforms for rapid design iteration. With advancing laboratory automation and DNA synthesis technologies, GenAI is positioned to achieve end-to-end optimization from element design to system construction within the next decade, driving transformative innovations across precision medicine, sustainable biotechnology, and biological manufacturing.

## RNA DESIGN

RNA molecules are fundamental to life, playing pivotal roles in gene regulation, catalysis, and the transfer of genetic information.<sup>279–281</sup> The integration of AI into RNA research has marked the beginning of a transformative era,<sup>282–283</sup> driving remarkable progress in analyzing RNA sequences, structures, and functions.<sup>284–286</sup> This synergy is revolutionizing our understanding of RNA's diverse roles and accelerating the design of RNA molecules with tailored complexity.

While early AI applications focused on predicting RNA properties, the field is undergoing a significant paradigm shift towards generative approaches. This evolution is crucial, as GenAI offers the ability to design entirely novel RNA molecules *de novo*, rather than merely optimizing existing ones. Although this area is still maturing, GenAI has already demonstrated enormous potential in tasks such as RNA inverse folding and functional sequence optimization,<sup>129,133,287–289</sup> opening new avenues for creating novel RNA therapeutics, gene therapy vectors, and synthetic biology tools.<sup>140,290–291</sup> Figure 3B outlines the two pillars of AI-driven RNA design: sequence-level tasks, such as RNA Inverse Folding and mRNA optimization, and structure-level challenges, spanning from prediction to structure design. This section examines the application of GenAI in RNA design, tracing the evolution from predictive methods to the forefront of generative models, and discussing the significant clinical and ethical implications of this technology.

## Generative design of RNA sequences

**RNA inverse folding.** RNA inverse folding presents a fundamental challenge in computational biology: navigating the vast sequence space to identify candidates that satisfy both structural and functional constraints. Traditional methods relying on thermodynamic models and heuristic optimization, such as adaptive walks in the Vienna RNA package<sup>292</sup> and stochastic search in INFO-RNA,<sup>293</sup> established the foundation but suffer from limited explo-

ration of sequence space and susceptibility to local optima. The emergence of GenAI has fundamentally shifted this landscape, introducing distinct methodological approaches that differ not only in their algorithms but also in their fundamental assumptions about how structure-sequence relationships should be learned and applied. Early machine learning attempts, such as RNAsynth, utilized constraint learning approaches.<sup>289</sup> However, contemporary GenAI methods exhibit fundamental differences in their design principles across multiple key dimensions.

First, approaches vary significantly in how they represent RNA structure: while methods operating in 2D secondary structure space (like constraint-based frameworks) offer computational efficiency, geometric deep learning approaches such as gRNAde process full 3D complexity through graph neural networks,<sup>129</sup> providing better handling of long-range interactions but requiring higher computational resources and quality structural data. Second, learning strategies exhibit a clear trade-off between pure data-driven approaches, such as RhoDesign, which utilize large-scale structural datasets for end-to-end learning,<sup>130</sup> and physics-informed methods that incorporate biological constraints—the former may suffer from dataset biases. At the same time, the latter faces increased model complexity. Third, generation approaches differ in their optimization strategy: single-step mapping methods prioritize computational speed, while iterative refinement methods, including RL frameworks<sup>131–132</sup> and DMs like RiboDiffusion,<sup>133</sup> enable more thorough sequence space exploration through multi-step optimization but face challenges in training stability and computational cost. These GenAI advances have enabled significant progress in designing complex riboswitches and RNA sensors,<sup>294–295</sup> yet important challenges remain in handling complex tertiary interactions, particularly pseudoknots, and ensuring biological functionality beyond structural conformity. Future progress will likely require hybrid approaches that combine the strengths of different methods while addressing their respective limitations through better integration of biological knowledge, structural information, and generative modeling techniques.

**mRNA sequence optimization for therapeutics.** mRNA sequence optimization has become a key frontier in biotechnology, with applications ranging from enhancing protein expression to developing mRNA-based therapeutics.<sup>140,290–291</sup> Initially, progress was driven by predictive models that aimed to understand existing relationships between sequences and functions. Machine learning algorithms revealed key determinants of translation efficiency,<sup>296</sup> while deep learning architectures optimized specific components like 5'UTRs for improved mRNA delivery and ribosome loading.<sup>297</sup> These approaches also revolutionized coding sequence optimization, surpassing traditional codon optimization techniques<sup>298</sup> and enabling direct prediction of protein expression levels from nucleotide sequences.<sup>299</sup> Integrated algorithms can now optimize multiple mRNA components simultaneously based on target protein sequences.<sup>300</sup> However, the inherent limitation of predictive approaches—their dependence on existing sequence variations—has driven the field toward generative methods capable of creating entirely novel mRNA designs.

The transition to generative mRNA design has revealed distinct methodological approaches that differ in their core design strategies. Data-efficient methods focus on leveraging evolutionary patterns within RNA families, such as approaches that integrate alignment and secondary structure information.<sup>301</sup> In contrast, comprehensive generative frameworks, like GEMORNA, attempt to design both coding and untranslated regions simultaneously for enhanced translation capabilities.<sup>134</sup> Foundational language models represent another strategy, with approaches like GenerRNA providing broad pre-training followed by task-specific fine-tuning,<sup>135</sup> which contrasts with direct translation methods like RNATranslator, which frame mRNA design as protein-to-RNA sequence mapping.<sup>136</sup> Recent innovations have also introduced novel generation architectures, including biologically inspired bidirectional approaches in RNA-BANG<sup>137</sup> and motif scaffolding through masked DMs in EvoFlow-RNA.<sup>138</sup> Flow-matching methods, such as RNACG, offer alternative generative paradigms that combine sequence generation with attribute prediction.<sup>139</sup> While these advances have significantly enhanced our ability to design mRNA with desired expression profiles, stability, and regulatory properties,<sup>140</sup> challenges remain in balancing multiple competing objectives and ensuring reliable translation of computational predictions to

experimental performance. The field's continued progress will depend on the better integration of biological constraints with generative capabilities and the development of more robust validation frameworks for designed sequences.

### From structure prediction to generative design

RNA structure design represents a rapidly evolving frontier where AI applications are transitioning from traditional predictive approaches toward generative design capabilities. Traditional approaches for predicting RNA secondary and tertiary structures have evolved from thermodynamic models, such as mFold<sup>302</sup> and UNAFold,<sup>303</sup> to sophisticated AI-driven methods that significantly improve accuracy and handle complex structural features. Deep learning has transformed both secondary structure prediction through methods like MXfold2<sup>304</sup> and UFold,<sup>305</sup> which can handle complex motifs including pseudoknots, and tertiary structure prediction via approaches like RhoFold<sup>306</sup> that achieve sub-4 Å accuracy by adapting insights from protein folding breakthroughs. However, these predictive advances, while crucial for understanding existing RNA structures, fundamentally address the forward problem of determining what structure a given sequence will form. The ultimate challenge, however, lies in moving beyond prediction to the generative design of entirely novel RNA structures with desired functional properties—a capability that requires generative approaches capable of creating unprecedented structural architectures.

The emerging landscape of generative RNA structure design reveals distinct methodological approaches that differ fundamentally in their design philosophy and technical implementation. Direct generative approaches like RNAgrail employ DMs for 3D structure prediction,<sup>307</sup> while integrated co-design methods such as RNAFlow leverage flow matching to simultaneously generate sequences and structures, particularly for conditional tasks like protein-binding aptamer design.<sup>308</sup> RNAFlow's innovation lies in integrating pre-trained structure prediction networks (RoseTTAFold2NA<sup>141</sup>) with inverse folding models within its generative loop, enabling design without costly fine-tuning while capturing RNA's conformational flexibility—a critical aspect often overlooked in static design approaches. Alternatively, hybrid strategies exemplified by Transformer-based sgRNA library generation followed by AlphaFold 3's structural filtering demonstrate how generative sequence models can be effectively combined with cutting-edge structure prediction tools for high-throughput functional screening.<sup>309</sup> Each approach faces distinct challenges: direct generative methods must ensure the physical feasibility of designed structures, co-design methods must balance competing sequence and structural objectives, while hybrid approaches require careful integration of disparate model architectures. As the field matures, the critical challenge will be moving beyond static structural design toward creating dynamic, functionally adaptive RNA molecules that can respond to cellular environments—a capability that will likely require novel generative frameworks integrating structural prediction, conformational sampling, and functional optimization within unified design pipelines.

### Summary

GenAI methods are beginning to make significant strides in RNA design, with clear promise in RNA inverse folding and sequence optimization. However, the field is still in its infancy. A key challenge lies in bridging the gap between sequence, structure, and function. The limited availability of high-quality experimental data hinders the development of robust generative models, and the experimental validation of designed RNAs remains a bottleneck. Future progress will depend on developing generative models that can design molecules based on functional requirements, not just structural templates.

The translation of these AI-designed molecules into clinical applications, especially for RNA therapeutics and vaccines, is a critical goal. The rapid development of COVID-19 vaccines highlighted the potential of computational approaches,<sup>310</sup> and the next wave of therapies for cancer and rare genetic diseases will increasingly rely on AI to co-optimize for potency, stability, and safety.<sup>311</sup> However, the successful clinical adoption of these advanced therapies hinges on overcoming significant biological barriers, such as targeted delivery and minimizing immunogenicity,<sup>312–313</sup> which are themselves complex design problems for AI to solve.

Finally, the increasing power of GenAI in biology necessitates urgent and

thoughtful consideration of the associated ethical and security implications. The use of patient data to train models requires robust privacy frameworks,<sup>314</sup> and developers must actively work to mitigate algorithmic biases to ensure that new therapies do not widen existing health disparities.<sup>315</sup> Furthermore, the "dual-use" nature of this technology—where tools for designing therapeutics could be misused to create biothreats—poses a serious security risk.<sup>316</sup> Proactive governance, international collaboration, and a culture of responsibility within the scientific community are essential to navigate these challenges and ensure that the transformative power of AI-driven RNA design is harnessed safely and equitably for the benefit of humanity.

### PROTEIN DESIGN

Proteins are the "machines of life" within organisms, playing a crucial role in nearly all life processes such as transcription, translation, signal transduction, and cell cycle regulation.<sup>317–319</sup> Natural proteins encode their amino acid sequences in a highly economical and precise manner, and they spontaneously fold into specific three-dimensional structures to achieve their biological activity. Exploring the relationship between protein structure and function has been a focus of fundamental medical and biological research over the past few decades.<sup>320–324</sup> As research into protein function deepens and practical applications expand, natural proteins are no longer able to meet the growing demands of humanity. The modification and design of proteins have shifted from relying on random mutations and directed evolution of natural proteins to rational design and even *de novo* design of entirely new proteins with specific functions.<sup>325–327</sup> In the field of protein design, recent years have seen a gradual shift in design methods from scoring functions based on physicochemical principles to strategies utilizing deep learning.<sup>325,328–333</sup> Based on the input and output modalities, these deep learning approaches can be broadly classified into four paradigms: sequence-only design, inverse folding, backbone design, and sequence-structure co-design, as summarized in Figure 3C. In this section, we review state-of-the-art deep learning approaches for protein design, categorized into three areas based on specific protein types: (A) antibody design; (B) peptide design; and (C) enzyme design.

#### Antibody design

Antibodies, also known as immunoglobulins (Ig), are essential proteins produced by B lymphocytes in response to antigens or haptens. They play a crucial role in the immune system by reversibly and specifically binding to corresponding antigens to form antigen-antibody complexes, enabling the recognition of foreign or self-antigens during adaptive immune responses. In particular, monoclonal antibodies (mAbs) are highly homogeneous antibodies derived from single B lymphocyte or hybridoma cell clones, targeting specific antigenic epitopes. These antibodies, predominantly based on the IgG framework, have high specificity, long half-life, and minimal off-target toxicity, making them valuable as therapeutic macromolecules. The development of monoclonal antibodies has become a cornerstone in therapeutic research, especially due to their broad market potential and diverse clinical applications.

Traditional methods for developing antibodies are labor-intensive and costly, which restricts their scalability and broader application. To address this, AI techniques are being increasingly adopted to optimize existing antibodies and generate new ones targeting specific epitopes, combining computational approaches with wet-lab techniques. Recent advances include sequence-based and structural approaches for antibody optimization and design. For instance, Hie et al.<sup>142</sup> employed a protein language model (ESM-1v) to introduce mutations improving the affinities of antibodies targeting pathogens like influenza A, Ebola virus, and coronaviruses. Similarly, Bennett et al.<sup>143</sup> fine-tuned the RFdiffusion model using thousands of experimentally determined antigen-antibody complexes, enabling the design of thousands of antibodies, including nanobodies, against diverse bacterial and viral antigens with atomic-level precision.

Monoclonal antibodies have shown remarkable potential in treating cancer and infectious diseases. Anti-tumor monoclonal antibodies are designed to target specific proteins involved in tumor growth and immune evasion, such as Her2 and PD-L1, forming the basis of several cancer therapies. Against pathogens, antibodies are engineered to recognize and neutralize key anti-

genic sites, such as the hemagglutinin of influenza viruses, the spike protein RBD of coronaviruses, and the F protein of RSV. We introduce various AI techniques proposed for antibody design against tumors and pathogens as follows.

**Antibody against tumor.** Monoclonal antibody drugs for cancer were first approved for marketing (the FDA approved Rituximab on November 26, 1997, and it was the world's first marketed anti-tumor monoclonal antibody), which has always been a hot spot in the development of monoclonal antibody drugs. According to their different targets, antibodies can be divided into antibodies targeting tumor-related molecules (such as VEGF, CD20, HER2 and EGFR) and antibodies targeting immune checkpoint inhibitors (such as PD-L1 and PD-1).

AU-007 is an anti-IL-2 antibody designed by Biologix and is in clinical trials.<sup>334–335</sup> Based on the known structural information of the IL-2 trimer receptor complex, Biologix first used machine learning algorithms to screen for reasonable template antibodies in the antigen-antibody complex database. Although the template antibody cannot bind to IL-2, the machine learning algorithm predicts that it is complementary to the surface structure of IL-2, suggesting that the template antibody has the potential for further development. Next, Biologix used a machine learning model to recommend key residue sites, and employed a focused library to train a point mutation prediction model to obtain a recommended mutation combination consisting of three mutations. Finally, Biologix used multiple machine learning models for antibody humanization and screening for stability and yield.

Parkinson et al.<sup>144</sup> developed a sequence-label model to predict yeast display affinities based on sequence embeddings pretrained on native antibody sequences. The model was then used to optimize atezolizumab for binding to programmed cell death ligand 1 (PD-L1). Mason et al.<sup>145</sup> developed a discriminative sequence-label CNN by training on deep mutational scan labels. It simultaneously leads to a marginal improvement in trastuzumab's affinity for HER2 while also considering viscosity, solubility, and immunogenicity. Makowski et al.<sup>146</sup> improved binding to the kinase c-MET by training a discriminative sequence-label model on a mutational scan of emibetuzumab.

These works mainly aim to enhance the features of existing antibodies. For specific antibodies, wet experiments are required to determine labels for training models, and phage-display selection and other techniques are used to screen candidate antibodies with high throughput. Some works are linked to the *de novo* generation of antibody sequences, such as Shanehsazzadeh et al.,<sup>147</sup> who generated new heavy-chain CDRs in a model conditioned on an existing antibody-antigen structure (trastuzumab-human HER2) and light-chain CDR sequences from a known antibody (trastuzumab).

**Antibody against pathogens.** Immune and infectious diseases are the second largest field of monoclonal antibody drug development. The pandemic of SARS-CoV-2 has made the research and design of SARS-CoV-2 neutralizing antibodies a popular research topic. Due to the continuous emergence of SARS-CoV-2 mutants, the design of corresponding binding and neutralizing antibodies has also become the cornerstone of many deep learning algorithms.

Shanker et al.<sup>148</sup> used a structure-informed protein language model to predict high-fitness sequences constrained by the known structure of the antibody or antibody-antigen complex. The authors screened about 30 variants of two therapeutic clinical antibodies used to treat severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection. Up to 25-fold improvement was achieved in neutralization and 37-fold improvement in affinity against antibody-escaped viral variants of concern BQ.1.1 and XBB.1.5.

Desautels et al.<sup>149</sup> designed a generative unconstrained intelligent engineering platform (GUIDE) - a combination of high-performance computing, molecular simulation, and deep learning, which can jointly optimize the binding affinity for multiple antigen targets, as well as other key properties (such as thermal stability). Within 3 weeks, the computing platform repaired the activity of the antibody COV2-2130 against the SARS-CoV-2 Omicron variant.

A pre-trained antibody generative LLM (PALM-H3) for the *de novo* generation of artificial antibodies heavy chain complementarity-determining region 3 (CDRH3) with desired binding specificity for SARS-CoV-2 antigens was proposed by He et al.<sup>150</sup> Additionally, the authors developed a high-precision model called A2Binder that predicts binding specificity and affinity by pairing

antigen epitope sequences with antibody sequences.

Due to the high mutation rates in SARS-CoV-2, antibodies could lose their binding affinity. Although not designed from scratch, it is also challenging to design a high-affinity antibody against the mutant strain from a maternal antibody that binds weakly or does not bind at all. The above work saves resources and labor by simulating calculations instead of a wet experiment. It is worth mentioning that Ig-VAE<sup>151</sup> was used to generate a single-domain binder to the ACE2 epitope of the SARS-CoV-2 receptor binding domain (RBD).

## Peptide design

Polypeptides are molecules formed from two or more amino acids linked by peptide bonds. In living organisms, peptides play a variety of key roles, such as hormones (e.g., insulin and growth hormone), enzymes, antibodies, and other biologically active molecules. They are involved in regulating many physiological and metabolic processes in living organisms, offering numerous benefits, including minimized side effects, high specificity, high efficacy, high bioactivity, and resistance to the development of drug resistance. To date, more than 100 peptides are on the market for the treatment of various diseases such as cancer, diabetes, osteoporosis, multiple sclerosis, HIV infection, and chronic pain. The traditional methods for designing and synthesizing peptides primarily include chemical synthesis and biosynthesis. Among these, chemical synthesis methods, such as Solid-Phase Peptide Synthesis (SPPS), have been widely employed. However, traditional methods still face some limitations in peptide design and synthesis, which require much time and economic costs.

AI has rapidly evolved in recent years, and computer-aided drug design incorporating AI algorithms can overcome the difficulties and barriers associated with conventional drug research and screening. The development of therapeutic peptides, particularly antimicrobial and anticancer peptides, utilizes AI algorithms, including deep learning and conventional machine learning-based algorithms. By applying AI tools to support peptide drug development, drug screening can be performed quickly and efficiently, thereby significantly reducing the time and cost associated with the process. Recent reviews on peptide discovery cover the principles of both traditional machine learning and AI-based approaches. Deep learning methods are being gradually applied to optimize existing peptides and peptides against specific targets, resulting in a new paradigm that combines computational methods with wet lab techniques.

**Antimicrobial peptide.** Antimicrobial peptides (AMPs) offer alternatives to conventional antibiotics as promising agents for anti-microbial drug resistance. AMPs are short amino acid sequences (usually between 6 and 50 residues in length) with an overall positive charge and a large number of hydrophobic amino acids that kill various bacteria, viruses and fungi by disrupting the membranes, specific target binding, immunomodulation, anti-biofilm activity, and interference with metabolic processes to kill a variety of bacteria, viruses, and fungi.<sup>336–338</sup> AMPs have a variety of properties including activity, toxicity, stability, and synergy.

AI models have been developed to predict the properties of AMP discovery and development, including antimicrobial activity, toxicity, stability, cell penetration, and mechanism of action. AMP classification is a method to broadly categorize peptides as AMPs or non-AMPs. For example, AMPlify,<sup>152</sup> which is a bi-directional Long Short-Term Memory (LSTM) model with an additional mechanism of multiple heads of attention. AMP-BERT,<sup>153</sup> a bi-directional encoder representation of protein sequences model pre-trained on protein sequences, uses mainly deep learning to automatically obtain descriptive features for classification.

Depending on the construction of the dataset, antimicrobial activity prediction methods can provide predictions for specific genera,<sup>154, 155</sup> species<sup>157</sup> or strains.<sup>158</sup> More generally, microbe-specific strain (MSS) prediction methods<sup>155, 158</sup> predict the activity of a given peptide bacterial strain pair. These methods utilize genomic information about the strains, including similarities between strains and oligonucleotide composition. An activity comparison method was introduced by Losin and Veltri,<sup>154</sup> who devised a model employing a Siamese neural network, which predicts the MIC difference for each antimicrobial peptide pair. There are also many proposed toxic/nontoxic or hemolytic/nonhemolytic classifiers for peptides.<sup>156, 159–161</sup> In a recent study, Salem et al.<sup>159</sup> used transfer learning to first identify secreted peptides using a



LLM and then fine-tuned it for hemolytic activity classification.

One traditional approach to discovering AMPs is to scan a given peptide database using a discriminator. Similar approaches are still being explored today, e.g., through exhaustive screening of large peptide libraries, or by mining peptides from macrogenomic data. Alternatively, recent studies aim to discover AMPs using GenAI. In unconstrained generation, peptides are freely sampled from an ab initio model. In contrast, analog generation generates analogs using a given peptide as a prototype. Analog generation can be performed in single or multiple steps, with more steps producing peptides that may be more optimized but less similar to the prototype. In order to obtain peptides with desired AMP properties, it is necessary to control the generation process based on these properties. Controllable properties of peptides include AMP, activity, toxicity, microbial target, target mechanism, hydrophobicity, secondary structure, and sequence length. Controlled AMP generation was obtained by discriminator-guided filtering, pure learning, latent space sampling, conditional generation, and optimized generation.

Conditional generation models are equipped with additional variables corresponding to conditions such as selected AMP properties and are trained to generate samples that satisfy these conditions. Conditional generators include GAN-based<sup>162, 163</sup> and GNN<sup>164</sup> models. We recently proposed HydrAMP,<sup>165</sup> a conditional VAE model optimized for both unconstrained and analog generation of AMPs. HydrAMP uniquely enables temperature-controlled exploration of peptide variants, generating analogs with enhanced activity (e.g., Hydraganan-1 with MIC = 2  $\mu\text{g/mL}$  against *E. coli*) or inactive derivatives from non-AMP prototypes. Wet-lab validation confirmed that 75% of selected peptides exhibited low toxicity (HC50  $\geq 512 \mu\text{g/mL}$ ), while maintaining potent activity against drug-resistant strains, demonstrating its utility in rational AMP design. Das et al.<sup>166</sup> developed a conditional latent space sampling (CLaSS) approach, where attribute classifiers guided the generation of AMPs under dual constraints of antibacterial activity and low toxicity, yielding potent candidates Y112 and FK13.

Optimized generative models aim to modify a given query peptide to improve its performance. AMP design employs state-of-the-art optimization algorithms such as Bayesian optimization<sup>168</sup> or multi-objective evolutionary algorithms.<sup>169</sup> However, emerging AI-driven approaches combine generative models with optimization steps. For example, Hoffman et al.<sup>170</sup> introduced a zero-order optimization model based on gradient descent to convert toxic peptides into non-toxic peptides while maintaining antimicrobial properties. The method achieved rapid success within just two iterative cycles: experimentally validated peptides (e.g., 1-7 and 2-5) demonstrated enhanced antimicrobial activity (with MIC values as low as 1.56  $\mu\text{M}$  against *E. coli* DH5 $\alpha$ ) and minimal hemolytic toxicity ( $>200 \mu\text{M}$  HC50). In a binary VAE,<sup>171</sup> the distance of each peptide from the Pareto front was scored by non-dominant ordering, and then the predicted scores were optimized by quantum annealing. As a proof-of-concept, the model was used to optimize easily computable properties such as charge, density, instability index, and Borman index. Finally, Jain et al.<sup>167</sup> proposed an active learning algorithm utilizing cognitive uncertainty estimation and GFlowNets as a generator of multiple candidate active peptides.

**Antioxidant peptide.** Low concentrations of free radicals play a crucial role in various human biological functions, including apoptosis and protein phosphorylation processes. However, excess free radicals in the body are often associated with diseases such as Alzheimer's disease, cancer and other chronic diseases. Therefore, it is crucial to discover safe, efficient and readily available antioxidants to prevent the harmful effects of free radicals on cells or to mitigate their negative effects. Antioxidant peptides are protein fragments with specialized structures that help fight oxidative stress in the body. Antioxidant peptides are a class of natural bioactive peptides that have been widely studied in recent years. As a kind of natural antioxidant with relatively simple structure, easy absorption, good stability and no immunoreactivity, antioxidant peptides not only have strong antioxidant activity, but also other health benefits such as lowering blood pressure and anticancer, which are attracting more and more attention in the field of food and medicinal nutraceuticals.

In quantitative structure-activity relationship (QSAR) studies of peptide bioactivity, machine learning techniques have been effectively applied for feature selection and model development. Deep learning-based QSAR tools

AnOxPePred and AnOxPP were developed for the prediction and design of antioxidant peptides. AOPAnOxPePred<sup>172</sup> is a method for predicting the antioxidant properties of peptides using deep learning. The tool combines various machine learning and deep learning algorithms to achieve efficient identification of antioxidant peptides by feature extraction and model training on protein sequences. AnOxPP,<sup>173</sup> on the other hand, predicts antioxidant peptides using bi-directional LSTM neural networks.

Unconstrained generative machine learning models provide an alternative to traditional methods for the rapid discovery of antioxidant peptides with desired functionality. Hesamzadeh et al.<sup>174</sup> developed the first generative model based on recurrent neural networks (RNNs) for the ab initio design of novel AOPs. Experimental validation through DPPH and hemolysis assays confirmed that peptides like GP12 exhibit potent antioxidant activity without hemolytic effects. Molecular dynamics (MD) simulations further elucidated their interaction mechanisms with the Keap1 protein.

**Anticancer peptide.** Anti-cancer peptides (ACPs) are a class of short peptides with anti-tumor activity, usually consisting of 2-50 amino acid residues, with a relative molecular mass of 50-5,000. In terms of amino acid composition, ACPs are rich in positively charged lysine, arginine, and histidine, as well as hydrophobic leucine, which can endow ACPs with cationic properties and hydrophobicity. ACPs exhibit a variety of anti-cancer properties, such as inducing apoptosis of cancer cells, inhibiting angiogenesis, and preventing cancer cell proliferation and migration. Anticancer peptides work through various mechanisms, including binding to cancer cell membrane receptors or intracellular targets, activating apoptotic pathways, and disrupting the metabolic and growth environments of cancer cells, thus realizing effective killing of cancer cells.

By analyzing the amino acid sequence of a specific peptide to predict whether it has anticancer properties, AntiCP,<sup>175</sup> as the first computational method for identifying ACP, employs three hand-crafted features in a support vector machine (SVM) model. The emergence of deep learning techniques has greatly improved the accuracy and efficiency of prediction. For example, ACP-DL<sup>176</sup> uses a LSTM model to extract features from peptide sequences. ACPred-LAF<sup>177</sup> developed a deep learning method based on a transformer encoder with learnable and adaptive embedding.

An equally critical aspect is the prediction of which cancer types ACPs can effectively target. These predictions are critical for customizing the use of ACPs in cancer treatment, allowing therapies to be specifically designed based on the unique characteristics of the cancer. xDeep-AcPEP<sup>178</sup> used multitask learning in a CNN framework to predict IC50 values for ACPs in six different tissues. xACP-MLC<sup>179</sup> used a two-tiered prediction model, including one for identification and another for type classification. Each model is trained on a different dataset but shares the same architecture. ACPScanner<sup>180</sup> proposed a two-level prediction framework that utilizes pre-trained protein models, protein secondary structure predictions, and physico-chemical properties to represent features. LightGBM and GAT algorithms are employed to learn from these features and subsequently combine their results to improve prediction accuracy. DUO-ACP<sup>181</sup> utilizes pre-trained protein models to generate amino acid features, achieving the best results to date on this multi-label classification problem.

In one study, an RNN model of a LSTM unit was trained with 10,000 sequences of amphiphilic peptides assumed to be  $\alpha$ -helical cations, and then fine-tuned on 26 ACPs known to have anticancer activity through a transfer learning approach.<sup>182</sup> The optimized model generated 1,000 novel amino acid sequences, 98% of which were unique. Of the 12 peptides tested synthetically, 10 exhibited activity against breast cancer MCF7 cells and 6 were able to kill cancer cells highly selectively without affecting human erythrocytes. This study demonstrated the ability of deep generative modeling to implicitly extract design rules from known ACPs and design completely new ACPs from scratch without explicit structure-activity relationship information.

Another study proposed a GAN-based GANDALF peptide design system aimed at generating peptides against specific protein targets.<sup>183</sup> Unlike existing GAN methods that can only generate small molecules, GANDALF uses two networks to generate peptide sequences and structures separately, and incorporates active atoms and other data to generate complete peptide structures and predict peptide binding affinities to targets. In this study, a variety of peptides were generated for three targets, PD-1, PDL-1 and CTLA-4,

and the best peptides generated were found to have comparable binding affinities and three-dimensional binding aptitude to the FDA-approved drugs, as well as unique sequences. This work demonstrates the potential of the deep generation model in designing ACPs with favorable bioactivity and structural features.

### Enzyme design

Enzyme engineering and *de novo* enzyme design have emerged as breakthrough research domains, transforming our understanding and application of biomolecular functions.<sup>339-341</sup> In particular, various AI techniques have been applied not only to enhance the functionality of existing enzymes more efficiently but also to design novel enzymes with specific catalytic activities.<sup>342-344</sup> According to the International Enzyme Commission (IEC), enzymes can be classified into seven major categories based on the nature of their catalytic reactions,<sup>345</sup> namely, oxidoreductases, transferases, hydrolases, lyases, isomerases, ligases, and translocases. It is worth noting that certain enzyme classes remain underexplored, with no AI techniques yet developed for their design or optimization. In the following, we provide a brief overview of AI-driven approaches successfully applied to specific enzyme classes, including oxidoreductases, hydrolases, and lyases.

**Oxidoreductases.** In recent years, there have been some advances in the *de novo* design of enzymes with specific functions, though successful cases remain rare. For example, the Baker group designed a beta-barrel protein<sup>184</sup> to achieve the luminescent function of green fluorescent protein (GFP). However, this design still requires manual insertion of chromophore molecules into the barrel structure to achieve luminescence, and the luminescence intensity is weaker than that of wild-type GFP. William P. Russ et al.<sup>185</sup> used the Potts model from statistical physics to learn the co-evolutionary information of amino acids at different sites in multiple sequence alignments (MSA) and combined this with Monte Carlo sampling to generate a batch of catalytic branched-chain isomerases. In 2021, Donatas Repecka et al.<sup>81</sup> utilized a generative adversarial neural network, ProteinGAN, to design and screen malate dehydrogenases with up to more than 100 site mutations, demonstrating catalytic functions similar to those of the wild-type protein. Approximately 24% of all designed sequences exhibited good solubility and biocatalytic activity. The Kemp elimination reaction, which converts benzisoxazole to salicylonitrile, is a crucial reaction for studying carbon-proton transfer. AI-assisted screening completed the directed evolution of the Kemp elimination enzyme in the laboratory.<sup>346</sup> The evolved enzyme HG3.17 is the most efficient Kemp elimination enzyme, with a catalytic efficiency 1000 times higher than the initially designed HG3, reaching the level of natural enzymes. This high efficiency stems from the effective synergy between the dual catalytic groups Asp127 and Gln50, which stabilize the negative charge of the reaction intermediate and optimize substrate binding at the active site. Crystallographic and NMR spectroscopy studies revealed that during evolution, HG3.17 adjusted protein backbone conformations, reduced non-productive conformations, and enhanced catalytic efficiency. These findings suggest that by improving design protocols and explicitly sampling backbone conformations, more efficient catalyst sequences can be predicted. A milestone study by the Baker group in 2023 designed an artificial luciferase.<sup>186</sup> They developed a deep learning-based family-wide hallucination method to generate a large number of protein structures with diverse pocket shapes, thereby designing luciferases capable of catalyzing the chemiluminescent oxidation of the luciferin substrates diphenylterazine and 2-deoxycoelenterazine. One small, thermally stable enzyme exhibited catalytic efficiency towards DTZ nearly equivalent to natural luciferase, but with significantly improved substrate specificity. Another notable example is the AI-guided discovery of an alternative oxidoreductase from *Sphingobium* sp. CSO (SsCSO), which catalyzes the transformation of 4-vinylguaiacol to vanillin. Identified using the CataPro platform, SsCSO showed an activity 19.53 times higher than the initial candidate enzyme, and further optimization through computational mutagenesis led to a 3.34-fold increase in activity compared to the wild-type.<sup>347</sup>

**Hydrolases.** The Alper team<sup>190</sup> utilized the MutCompute, a structure-based CNN, to identify single-point mutations in 10 PET hydrolases, with eight mutations experimentally validated to enhance enzyme functionality. Through screening and combination, the team developed a five-site mutant

that can completely degrade most PET products within one week. Assisted by AI, enzyme engineering has become more cost-effective and time-efficient compared to traditional directed evolution. More recently, in a landmark study published in 2025, Baker and colleagues<sup>191</sup> computationally designed a series of novel serine hydrolases with entirely new protein folds. By integrating the generative diffusion model RFDiffusion with the active-site preorganization assessment tool PLACER, the team overcame key challenges in modeling multi-step catalytic mechanisms. The resulting synthetic enzymes exhibited catalytic efficiencies approaching those of natural hydrolases, representing a major step forward in the *de novo* creation of functional biocatalysts. Bian Wu's group also made significant advances in PET hydrolase engineering by applying a greedy accumulated strategy (GRAPE) to guide the rational design of high-performance PET-degrading enzymes. Their optimized variant, TurboPETase, achieved nearly complete degradation of industrial PET waste under high-solid conditions.<sup>187</sup> This work demonstrates the effectiveness of AI-guided iterative design in developing biocatalysts for plastic degradation. In parallel, Liang Hong and collaborators leveraged protein language models and generative diffusion frameworks such as CPDiffusion to design synthetic nucleases and hydrolase-like enzymes with programmable substrate preferences. Their work emphasizes the potential of large-scale sequence generation and conditioning to expand enzyme functionality across unexplored sequence space.<sup>188</sup>

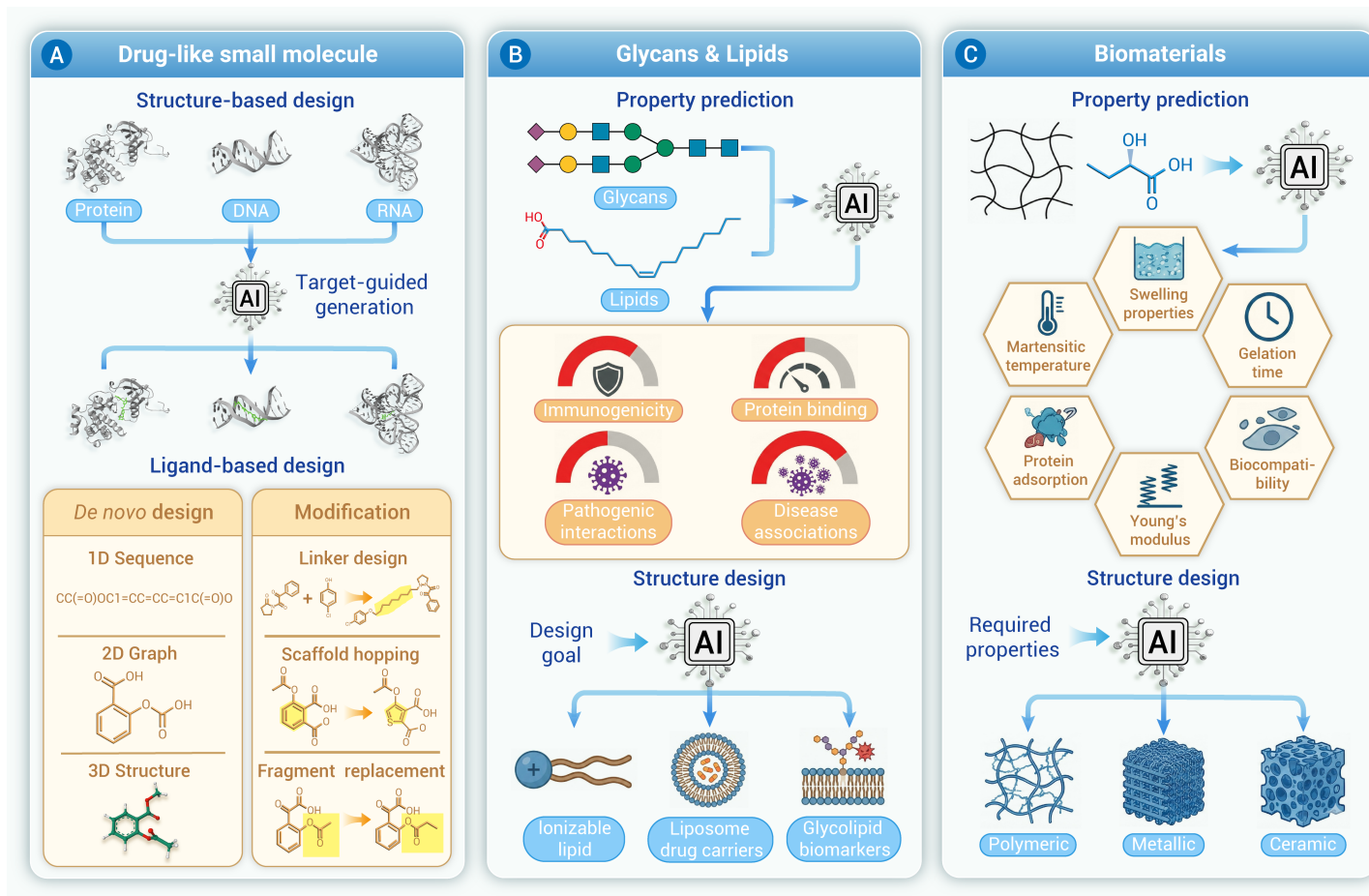
**Lyases.** Additionally, the combination of computational design and directed evolution has successfully developed efficient catalysts for bimolecular aldol condensation and Diels-Alder reactions.<sup>189</sup> RA95.5-8F is the most active aldolase to date, generated through ultra-high-throughput screening, with catalytic capabilities comparable to natural aldolases. During evolution, the catalytic site shifted from the initially designed Lys210 to the more reactive Lys83, forming a complex catalytic center that illustrates the role of the Lys-Tyr-Asn-Tyr tetrad and the designed hydrophobic pocket. This discovery highlights the powerful potential of combining computational design with experimental screening in the evolution of enzymes.

Overall, structure-based *de novo* protein design methods are innovative and novel. However, the success rate is low, often requiring high-throughput screening of vast design libraries to identify positive sequences, a process that is limited by both the demands of high-throughput experimentation and the constraints of oligonucleotide synthesis technology. This ultimately makes it challenging to design and synthesize protein sequences longer than 200 amino acids.

### Summary

Over the past few decades, research has focused on understanding the relationship between protein structure and function. With the increasing demand for proteins with novel functions, *de novo* protein design has gained significant attention. Deep learning has revolutionized protein design, with applications in antibody, peptide, and enzyme design. However, AI-driven protein generation faces significant challenges. AI-designed proteins may bind effectively to targets but fail to achieve the desired biological functions, and generative models sometimes produce non-natural structures that violate biophysical principles. Complex structures, such as transmembrane proteins, remain challenging to design, and some AI-generated proteins are difficult to synthesize using current techniques.

Beyond these technical hurdles, the clinical translation of AI-designed proteins presents distinct challenges. The rapid advancement of this technology raises significant biosafety concerns, as the capability to computationally generate novel biological agents could potentially be misused for harmful purposes.<sup>348</sup> At the therapeutic level, while these designed biologics show exceptional targeting capabilities for oncology and genetic disorders, their clinical validation faces unprecedented regulatory uncertainties. Current approval frameworks lack established guidelines for evaluating computationally generated proteins, particularly regarding long-term safety assessment and quality control standards for AI-predicted molecular properties. Furthermore, the specialized computational infrastructure and expertise required for developing, manufacturing, and monitoring these therapies may create significant accessibility barriers, potentially limiting their benefits to well-resourced healthcare systems.<sup>349</sup> Addressing these multifaceted challenges will require not only integrating robust biophysical constraints into AI models



**Figure 4. Design of biomolecules beyond the central dogma** (A) Drug-like small molecule design encompasses structure-based approaches guided by protein, DNA, or RNA targets, and ligand-based approaches including *de novo* design (1D sequence, 2D graph, 3D structure) and modification strategies (linker design, scaffold hopping, fragment replacement). (B) Glycans and lipids design involves property prediction (immunogenicity, protein binding, pathogenic interactions, disease associations) and structure design for ionizable lipids, liposome drug carriers, and glycolipid biomarkers. (C) Biomaterials design integrates property prediction (swelling, biocompatibility, protein adsorption, etc.) with structural design across polymeric, metallic, and ceramic materials.

but also developing comprehensive regulatory frameworks and ensuring equitable distribution mechanisms to realize the full therapeutic potential of this transformative technology.<sup>350</sup>

## DRUG-LIKE SMALL MOLECULE

This section specifically addresses the design of drug-like small molecules for therapeutic applications, distinguishing them from other molecular classes like glycans and lipids, which are discussed later. Drug-like small molecule design techniques aim to develop specific molecules that bind to biological targets while exhibiting desired properties. These techniques often rely on the structural information of biological molecules, such as DNA, RNA, or proteins. Once the structural information of a biological molecule has been obtained through techniques such as X-ray crystallography, NMR analysis, or generative algorithms, these biological targets can serve as a foundation for rational molecule design.<sup>351–352</sup>

From the methodology perspective, drug-like small molecule design can be categorized into three approaches: traditional small molecule design approaches, including high-throughput screening, molecular docking, virtual screening, and MD simulations, are often time-consuming and labor-intensive.<sup>353</sup> Beyond the traditional approaches, the advent of generative models has catalyzed explosive growth in small molecule design applications by implicitly searching larger chemical spaces.<sup>354</sup> Structure-based approaches analyze biological targets (DNAs, RNAs, proteins) to identify binding regions for lipophilic or polar groups,<sup>351</sup> with protein-guided design being the most common. Besides, ligand-based approaches design new molecules by analyzing the properties of known active compounds, specifically when target structures are unavailable. As illustrated in Figure 4A, these AI-driven strategies are broadly categorized into ligand-based approaches, encompassing *de*

*novo* design and modification, and structure-based approaches, which are guided by targets like proteins and nucleic acids. In the following section, we will focus on two GenAI-based approaches, i.e., structure-based and ligand-based, and then dive deeper into their characteristics.

### Structure-based drug-like small molecule design

Advances in structural biology enable detailed 3D analysis of biomolecule-ligand interactions, facilitating structure-based molecular design. Proteins are primary drug targets, including enzymes, receptors, ion channels, transporters, and kinases.<sup>355–359</sup> Nucleic acids, crucial for genetic processes and catalysis, represent additional therapeutic targets, offering potential treatments for viral diseases and cancers through disruption of gene expression.<sup>360</sup>

**Protein-guided drug-like small molecules design.** Most drugs exert their effects by interacting with specific protein targets. Understanding the complex characteristics of target proteins enables better design of drug-like molecules that can bind to them, thereby increasing the efficiency of drug discovery. Structure-based drug design utilizes information about the target protein to guide drug molecule design. It analyzes the structural features of proteins, such as the active sites and binding pockets, to design small molecule compounds that can tightly bind to and modulate their function. Protein-guided molecular design methods can be categorized into whole-structure-guided generation and binding site-guided generation based on the the specificity of the protein information utilized.

Whole structure-guided generation designs potential ligands by analyzing complete protein structures without requiring prior knowledge of specific binding sites. This method is suitable for cases where binding sites are unknown or difficult to identify, thereby expanding possibilities for drug

discovery. For instance, Grechishnikova et. al ingeniously utilized a Transformer model to "translate" protein sequences into small molecule sequences.<sup>192</sup> Alphadrag<sup>193</sup> builds upon the Transformer model, combining it with Monte Carlo tree search algorithms to generate molecules with high binding affinity. DeepTarget<sup>194</sup> employs a VAE to encode and decode molecular sequences for generation while using protein sequence representations as constraints to obtain molecules with high binding affinity. CDGCN<sup>195</sup> designed a conditional generation model based on graph convolutional networks, using the input protein sequence as a condition and generating novel molecular graphs through a sequential decoding scheme. PCMol<sup>196</sup> leverages protein embeddings from AlphaFold2 to guide a Transformer model in autoregressively generating SMILES strings for target-binding molecules. However, this approach faces two main challenges: processing complete protein structural data requires intensive computation, and the generated molecules may bind anywhere on the protein, requiring additional screening to determine likely binding modes.

Unlike whole structure-guided generation, binding site-guided generation methods utilize detailed information about specific binding sites on proteins to guide molecular design. With advancements in structural biology and computational chemistry, binding site-guided generation methods have evolved from simple surface matching to complex multi-scale modeling processes. Pocket surface-based generation methods use binding pocket geometry to guide molecular design. The core concept is to generate molecules that match the pocket shape, increasing protein-binding probability. LIGAN<sup>197</sup> uses GANs to generate 3D ligand shapes complementary to protein pockets, converting them to SMILES through a shape description network. DESERT<sup>198</sup> first sketches molecular shapes based on pocket geometry, then generates fitting molecules using a pre-trained generative model. While pocket surface shape-based generation methods effectively capture geometric features for spatial compatibility, they may overlook crucial chemical properties needed for strong binding affinity.

The atom or amino acid-based generation within the pocket is a more refined binding site-directed molecular design method. Compared to methods that only use surface shape information, this approach considers the specific types, positions, and chemical properties of atoms or amino acids in the binding site, thereby generating molecules with stronger interactions with the protein. EGCM<sup>199</sup> represents the 3D information of the pocket using the sorted eigenvalues of the coarse-grained Coulomb matrix (EGCM) of binding pocket atoms, then uses this as conditional input to a conditional recurrent neural network (cRNN) to generate molecules. GAT-VAE<sup>200</sup> constructs a residue graph of the pocket, learns structural features of the pocket through a graph attention model-based AE, then concatenates these with molecular features as interaction features, ultimately inputting them into a conditional decoder to generate target molecules. CProMG<sup>201</sup> separately encodes the atomic-level and amino acid-level structural information of the protein pocket and fuses them to obtain 3D features of the complex protein structure. It then utilizes multi-head interaction attention to calculate key interactions between small molecules and proteins, finally conditionally generating new molecules with specific properties in an autoregressive manner, conditioned on these interactions. Flag<sup>202</sup> is a conditional generation method that generates molecules one motif at a time, using the protein binding site as the condition for each iteration. Each iteration generates a new motif for the ligand molecule, eventually producing a complete ligand molecule. DrugGPS<sup>203</sup> considers that similar subpockets can often bind the same molecular fragments. It first constructs a global interaction graph of subpocket and molecular fragment nodes, using this as prior knowledge input to the model. It then iteratively generates molecular fragments along with connection sites and torsion angles between fragments, resulting in the 3D structure of the molecule.

These methods consider the structural information of the binding pocket, but they cannot determine the binding pose of the generated molecules within the pocket. With the development of geometric deep learning and DMs, an increasing number of methods can directly generate 3D coordinates of molecules. Targetdiff<sup>204</sup> achieves protein target-aware molecular design by using DMs, which learn a joint generation process for continuous atomic coordinates and discrete atom types through SE(3) equivariant networks. DecomDiff<sup>205</sup> divides molecules into scaffolds and arms based on the different roles of atoms in the molecule. Scaffolds are responsible for positioning the

molecule in the binding region, while arms contact the target to increase binding affinity. Based on this assumption, DecomDiff designs a heterogeneous graph diffusion model with a decomposition prior for arms and scaffolds, capable of simultaneously diffusing and generating atomic coordinates, atom types, and bond types. To better learn the binding patterns between proteins and small molecules, some methods explicitly utilize interaction information to generate small molecules. DeepICL<sup>16</sup> proposes an interaction-aware 3D molecule generation framework, achieving interaction-guided drug design by leveraging universal patterns of protein-ligand interactions as prior knowledge. NCIDiff,<sup>206</sup> based on a diffusion model, simultaneously denoises the non-covalent interaction types of protein-ligand edges and the 3D graph of ligand molecules. It emphasizes learning universal patterns of protein and ligand binding, generating ligands with more reliable NCIs.

A compelling case in point is the discovery of Rentosertib (INS018-055) by Insilico Medicine.<sup>361</sup> Their AI-driven workflow began with the identification of TNIK as a novel target for Idiopathic Pulmonary Fibrosis (IPF). Based on the TNIK protein structure, their generative chemistry platform, Chemistry42, then performed *de novo* design of novel inhibitors while simultaneously optimizing for binding affinity, selectivity, and pharmacokinetic properties. This process yielded the clinical candidate Rentosertib, which has since progressed to Phase II trials. This achievement powerfully validates that structure-based GenAI can effectively translate a novel target into a clinical-stage therapeutic, demonstrating its capability to efficiently produce drug candidates of significant clinical value.

**Nucleic acid-guided drug-like small molecules design.** Nucleic acids have been confirmed as molecular targets for numerous clinical anticancer chemotherapeutics.<sup>360</sup> Targeting nucleic acids with drug-like small molecules offers opportunities to modulate various cellular processes, particularly those associated with "undruggable" protein targets.<sup>362</sup> However, unlike proteins, nucleic acid targets, including RNAs and DNAs, have received much less attention in the structure-based drug-like small molecule design field. This limited attention is likely due to the challenges associated with the specific recognition of nucleic acids by small molecules and the lack of structural information on nucleic acid-small molecule complexes.<sup>360</sup> As mentioned earlier, over the past decades, various methods have been developed to predict and design complex nucleic acid structures. These methods have laid the groundwork for nucleic acid-based small molecule design.

RNAs, traditionally viewed as an intermediary between DNAs and proteins, have been extensively validated as a therapeutic target.<sup>363</sup> Functional RNA forms structured binding pockets for small molecules, as demonstrated by riboswitches that small molecule ligands can selectively bind to control gene expression.<sup>365</sup> Many antibiotics bind directly to ribosomal RNA through this mechanism. While traditional design methods include fluorescence-based and FRET-based approaches,<sup>363</sup> AI-driven design of RNA-targeted small molecules has been limited to virtual screening and binding prediction. DNA-targeted small molecule design has been relatively sparse due to the limited availability of well-characterized DNA structures. Unlike proteins, DNAs typically lack structured binding pockets, with small molecules targeting DNA through intercalation, alkylation, and DNA crosslinking.<sup>364</sup> Through molecular docking and simulations, AI-based methods can predict small-molecule binding modes with DNAs.

Building on protein-targeted AI techniques, AI methods have been adapted for nucleic acid-targeted small molecule design. A key challenge in RNA targeting is the inaccuracy of scoring functions in predicting small molecule binding conformations.<sup>365</sup> To address this, various AI approaches have been developed, including RNAPosers<sup>206</sup> using random forest classifiers, and AnnapuRNA<sup>208</sup> employing neural networks for pose scoring. Haga et al.<sup>209</sup> applied GNN architectures to identify compounds targeting miR-21. In industry applications, Merck<sup>366</sup> developed an ML-informed RNA-focused library comprising approximately 3,800 compounds, resulting in higher hit rates compared to conventional libraries.

AI methods have also been used to predict the binding affinity of small molecules to nucleic acid targets. These methods not only provide binding affinity data but also offer insights into other aspects of the binding event, aiding in hit-to-molecular modification. Szulc et al.<sup>210</sup> used molecular fingerprints to characterize small molecules and assessed the affinity prediction performance of 15 ML algorithms, finding that CatBoost, extremely randomized



trees, and light gradient boosting models performed best in predicting ligand binding. Krishnan et al.<sup>211</sup> developed a multivariate linear regression model to predict the binding affinities of small molecules to six RNA subtypes and conducted Shapley additive explanations analysis to quantify feature contributions to model predictions.

Grimberg and colleagues advanced the exploration of chemical space for RNA targets by designing small molecules with greater potency and selectivity.<sup>212</sup> The group trained a Lasso regression model on a dataset containing molecular descriptors of 791 small molecules featuring the 2-phenylthiazole moiety and their binding affinities for hairpin 91. This model identified key features predictive of RNA binding affinity. They also employed a decision tree model to classify small molecules as binding or non-binding and determined the critical descriptors for RNA binding. Additionally, they used a CNN to analyze unique geometrical patterns in small molecule structures, based on their SMILES representation and molecular images. Insights from these approaches led the group to design ten molecules containing the 2-phenylthiazole moiety, four of which were identified as potent inhibitors of the RNA target.

### Ligand-based drug-like small molecule design

Ligand-based drug-like small molecule design can be categorized into two approaches: *de novo* design, which entails creating entirely new molecular structures to explore uncharted chemical space, or molecular modification, which focuses on modifying existing compounds to enhance their efficacy, selectivity, pharmacokinetics, and safety. In cases where the target structure is unknown or difficult to obtain, analyzing active molecules known to interact with the target can facilitate the design of new molecules with similar or improved activity, thereby enhancing the efficiency of drug discovery. Ligand-based molecular generation utilizes information from known active molecules to guide drug design. It can be divided into two main directions: structure-reasonable molecule generation and multi-objective controllable generation.

**De novo design.** Structure-reasonable molecule generation is a fundamental task in ligand-directed molecular design, with the goal of generating molecules that are chemically structurally sound and easily synthesizable. The challenge of this task lies in exploring a sufficiently large chemical space to discover novel molecular structures while ensuring the reasonableness of these structures.

1D molecular generation linearizes molecular structures into SMILES strings, a representation that naturally lends itself to the adoption of mature techniques from Natural Language Processing (NLP). Early methods predominantly employed RNNs and their variants.<sup>213</sup> However, the superior Transformer architecture has now become the mainstream choice, owing to its powerful ability to capture long-range dependencies. Within this framework, the application of LLMs is particularly noteworthy and has catalyzed a diverse array of generation strategies. For instance, TSMMG<sup>214</sup> employs a novel teacher-student paradigm, trained on massive text-molecule pair data, which enables the LLM to translate complex natural language prompts into target molecules directly. Other work has focused on enriching the 1D sequence with additional dimensions of information. 3DSMILES-GPT,<sup>215</sup> for example, unifies a molecule's 2D (SMILES) and 3D (atomic coordinates) information into a single sequence, autoregressively generating molecules with precise spatial conformations via a Transformer decoder. Diverging from end-to-end generation, stepwise strategies have also shown immense potential. Sc2Mol<sup>216</sup> ingeniously combines different architectures: it first generates a core molecular scaffold using a VAE. Then it employs a Transformer model to "decorate" this scaffold with functional groups, achieving a modular and fine-grained construction of molecular structures. Concurrently, DMs have been introduced to this domain for their high-quality generation capabilities. A representative model, TGM-DLM,<sup>217</sup> adopts a two-stage diffusion process: the first stage generates a draft SMILES string guided by the text, while the second stage uses a dedicated correction network to rectify invalid sequences, thereby significantly improving the chemical validity of the generated molecules.

2D molecular generation represents molecules as graphs, a format that more naturally reflects their actual topological structure. To achieve high-fidelity graph generation, the field has adapted and extended various classic

model architectures. Seminal early works include JT-VAE,<sup>218</sup> which operates on a tree of valid chemical substructures to ensure chemical validity; MolGAN,<sup>80</sup> which applies adversarial learning to the graph's adjacency and feature matrices; and autoregressive models like GraphAF<sup>219</sup> which construct molecular graphs through a sequence of ordered decisions. Following the tremendous success of the attention mechanism, its principles have been adapted for graph data. GraphXForm,<sup>220</sup> for example, iteratively adds atoms and bonds to a dynamically growing subgraph to construct the molecule. Currently, DMs have become a research hotspot in graph generation. DiGress<sup>221</sup> directly learns to reverse the noising of node and edge types on discrete graph data. To tackle more complex structures, HGLDM<sup>222</sup> introduces a hierarchical diffusion process that denoises synchronously across multiple levels of abstraction (e.g., atoms, graphs), substantially enhancing the model's ability to capture intricate architectures. Furthermore, the SyCo<sup>223</sup> model explores the use of cross-dimensional information: it first generates a coarse 2D structural sketch as a prior, learns its corresponding 3D geometry using an equivariant neural network, and finally leverages this valuable spatial information to guide the generation of a higher-fidelity 2D molecular graph.

3D structure generation, which directly constructs atomic coordinates in three-dimensional space, is one of the most challenging yet promising frontiers in drug design. DMs have emerged as the predominant technology for this task. For instance, EDM<sup>89</sup> an equivariant diffusion model, innovatively designs a specialized noising scheme for the multimodal features of molecules—discrete atomic types and continuous coordinates—and adjusts input feature scales to enhance training stability. To avoid complex modeling of different variables, GeoLDM<sup>224</sup> employs an encoder to map molecules into a geometric latent space composed of invariant scalars and equivariant vectors, where the diffusion process is then performed, simplifying the modeling challenge. A common pitfall of traditional DMs is generating atoms without considering chemical bonds, leading to structural inconsistencies. To address this, MolDiff<sup>225</sup> pioneers the co-diffusion of atoms and bonds, explicitly modeling their interdependencies to ensure the chemical rationality of the molecules generated. To incorporate more physical priors, MD3MD<sup>226</sup> decomposes molecules into multi-scale graphs and assigns different weights to inter-atomic interactions at various scales to guide the diffusion process, making the generation more constrained by physical laws.

Multi-objective controlled generation is a key task in LBDD, aiming to create molecules that simultaneously satisfy multiple design criteria, such as high activity, specific physicochemical properties, a favorable pharmacokinetic profile (Absorption, Distribution, Metabolism, Excretion, and Toxicity - ADMET), and low toxicity. In this context, some new methods focus on optimizing the decision-making process itself. For example, MLPS<sup>227</sup> does not seek a single optimal solution but instead learns a direct mapping from user preferences to the entire Pareto front, thereby providing decision-makers with a comprehensive and queryable set of optimal solutions. In terms of model implementation, MCMG<sup>228</sup> was a pioneer in using a conditional Transformer, fine-tuned with knowledge distillation and RL, to achieve multi-objective optimization for properties like quantitative estimate of drug-likeness (QED), synthetic accessibility (SA), and activity. To more precisely control synthetic feasibility, TANGO<sup>229</sup> designed a novel reward function that, in conjunction with a retrosynthesis prediction tool (Syntheseus), guides the model to generate molecules composed of a predefined set of building blocks, fundamentally enhancing synthesizability. SVAE<sup>230</sup> designed a semi-supervised VAE that combines RL with the MGA toxicity prediction model to generate low-toxicity molecules. Although preliminary exploration has been made in the low-toxic molecular generation model, they also face the challenge of complexity and uncertainty of toxicity prediction. ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) property-controlled small molecule generation is a crucial and emerging research direction in drug design. Although there are currently no mature generative models that directly control ADMET properties, this direction is of significant importance and has the potential to bring revolutionary changes to drug discovery.

**Molecular modification.** *De novo* design generates new molecular structures from scratch to explore novel chemical spaces for therapeutic agents. In contrast, molecular modification improves existing compounds by enhancing their efficacy, selectivity, and pharmacokinetics. The molecular modification process can be classified into three main subtasks: linker design

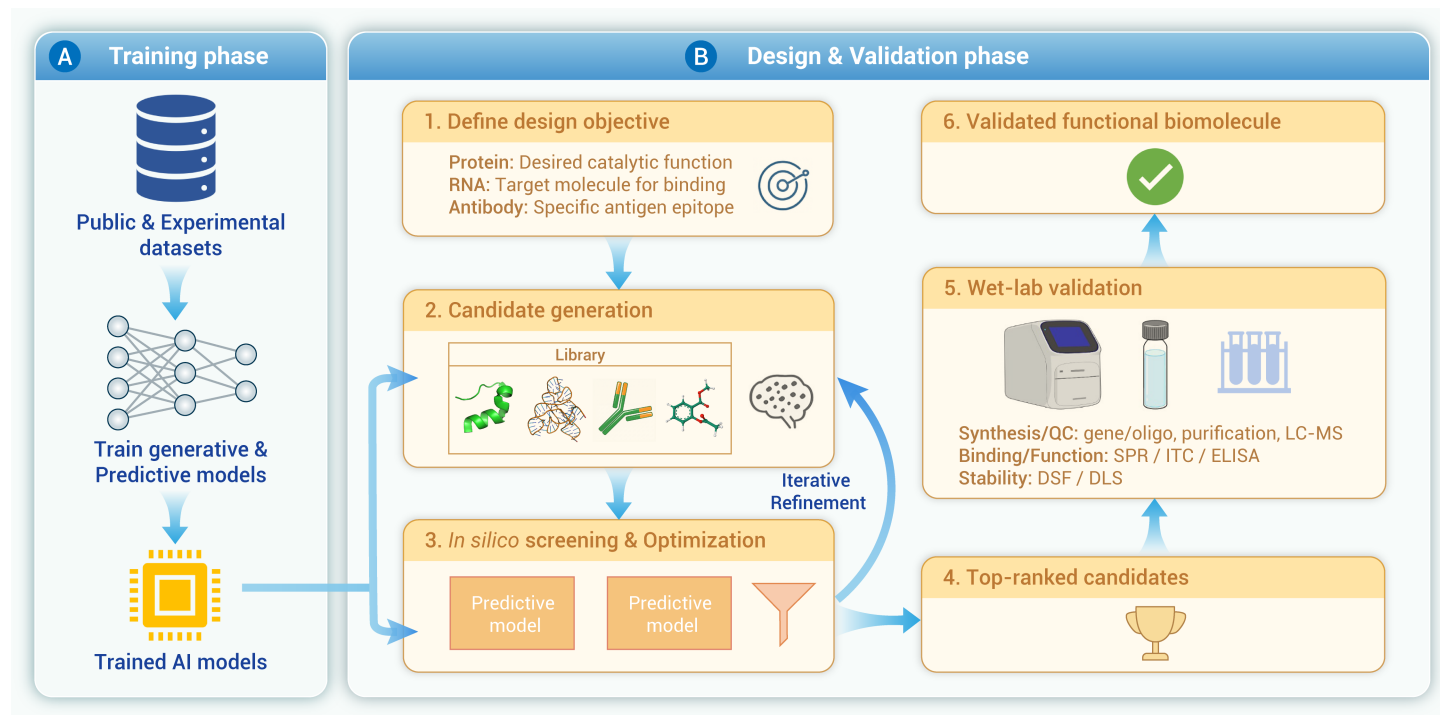


Figure 5. A general workflow for designing functional biomolecules using GenAI models.

for optimizing fragment connections, scaffold hopping for core structure replacement, and fragment replacement for improving binding interactions.<sup>367</sup>

Linker design is crucial in molecular modification, where low molecular weight fragments with weak binding affinity but strong interactions with protein pockets are connected.<sup>367</sup> Recent generative models have advanced this field significantly. DeLinker<sup>231</sup> uses a VAE to learn fragment and linker distributions, while SyntaLinker<sup>232</sup> employs a Transformer architecture to generate linkers by translating fragmented molecules into SMILES format. Its successor, SyntaLinker-Hybrid,<sup>233</sup> introduces fragment hybridization. 3DLinker<sup>234</sup> further advances the field by incorporating equivariant neural networks to consider geometric configurations during linker generation.

Scaffold hopping involves replacing the core structure, or scaffold, of a lead compound while retaining its functional activity. This strategy is employed to explore different chemical backbones that might offer better pharmacological profiles, reduce off-target effects, or bypass existing patent restrictions. Graph-GMVAE<sup>235</sup> enables scaffold sampling across three scales using a multivariate Gaussian distribution. DeepHop<sup>236</sup> treats scaffold hopping as a translation task, generating molecules with similar 3D structures but different 2D scaffolds. DiffHop<sup>237</sup> advances this field by incorporating interaction-awareness and improving generation precision in both structure and activity.

Fragment replacement involves the optimization of lead compounds by replacing or growing fragments to enhance binding affinity and fill unoccupied spaces within the protein pocket. DeepFrag<sup>238</sup> approaches this as a classification task, selecting fragments from a library. DEVELOP<sup>239</sup> uses a VAE to eliminate library dependency and incorporates CNN-extracted pharmacophore information. STRIFE<sup>240</sup> extends this by using fragment-based pharmacophore models. D3FG<sup>241</sup> advances the field through SO(3) diffusion architecture, predicting fragment positions while preserving physical properties like symmetry and rotation.

Additionally, some studies guide the modification of existing molecules by constraining specific properties to generate molecules with desired characteristics. Certain research focuses on searching for target molecules within the latent space learned by encoder-decoder models, guided by property predictors. JT-VAE<sup>218</sup> decomposes molecules into junction trees and uses gradient ascent in latent space to improve property scores. MoFlow,<sup>242</sup> a flow-based model, generates molecular bonds and similarly optimizes properties through gradient ascent. Alternative approaches use RL or genetic algorithms. GCPN<sup>243</sup> frames generation as a Markov decision process with validity and adversarial rewards, while GA-DNN<sup>244</sup> combines property scores and

discriminator scores in genetic algorithms to optimize molecules.

Currently, a targeted therapy for diffuse gastric cancer, SIGX1094, developed through AI screening and optimization, has entered Phase I clinical trials (NCT06739291). As a dual FAK/SRC inhibitor, SIGX1094 was rapidly identified from millions of compounds via an AI platform, with its chemical structure further optimized by AI models to enhance binding affinity. The FDA has granted it Orphan Drug Designation and Fast Track Designation. This milestone marks the critical transition of AI-driven molecular modification from laboratory research to clinical application.

### Summary

Drug-like small molecule design faces significant challenges despite recent advances<sup>201,204,215,217,225</sup> in AI-driven approaches. Although structure-based and ligand-based methods have shown promise, achieving optimal molecular properties remains complex due to the dynamic nature of biomolecular targets and limited high-quality training data. The field particularly struggles with simultaneously optimizing multiple parameters, including efficacy, ADMET properties, and synthesis feasibility. This challenge is evident in developing antibody-drug conjugates and fluorescent probes, where molecules must maintain both target specificity and desired physicochemical properties. Moreover, the gap between computational predictions and experimental validation creates a bottleneck in translating AI-generated candidates into practical applications. Current AI approaches also often lack interpretability in their decision-making process, limiting our ability to design drug-like molecules for novel targets or mechanisms rationally. These challenges underscore the importance of developing integrated approaches that combine AI with experimental validation and mechanistic understanding to advance drug-like small molecule design across therapeutic and diagnostic applications.

The design of functional biomolecules, from DNA and RNA to proteins and drug-like small molecules, follows a systematic AI-driven workflow that integrates computational prediction with experimental validation (Figure 5). This workflow consists of two main phases: a training phase that develops robust generative and predictive models from public and experimental datasets, and a design and validation phase that systematically progresses from defining design objectives through candidate generation, in silico screening, and wet-lab validation to produce functional biomolecules. This integrated approach provides a unified framework for addressing the unique challenges encountered across different biomolecular design applications.

## GLYCANS AND LIPIDS

Glycans and lipids are essential biomolecules that serve critical functions in biological systems. Glycans serve as the primary energy source for organisms, fueling various metabolic processes, including glycolysis and cellular respiration. In contrast, lipids serve as highly efficient energy storage molecules because of their high energy density. Beyond their roles in energy metabolism, lipids participate in various physiological processes, which include signal transduction, organ protection, and thermal insulation. Moreover, both glycans and lipids constitute fundamental components of cellular structures. While glycans help form cell walls in plants and bacteria, lipids serve as the main components of cell membranes to maintain their fluidity and stability. These essential functions demonstrate that glycans and lipids are indispensable for maintaining physiological balance and ensuring normal function in living organisms. However, natural glycans and lipids, biomacromolecules often face structural and functional limitations, which necessitate innovative design strategies to enhance their potential applications and efficiency.

To address the urgent need, AI has emerged as a powerful tool in glycoinformatics and lipidomics, enabling advanced structural analysis, property prediction, and molecular optimization. Despite its promise, the application of AI in the *de novo* design of glycans and lipids has been a significant challenge, largely due to limited data availability and the complexity of these biomolecules. However, recent breakthroughs are beginning to directly address these hurdles, with deep generative models emerging as a powerful new approach for the direct, bottom-up design of synthesizable lipids.

As illustrated in Figure 4B, this section provides an overview of current AI techniques for predicting the properties of glycans and lipids (including immunogenicity, protein binding, pathogenic interactions, and disease associations), as well as AI-driven structure optimization approaches (featuring ionizable lipid design, optimized liposome drug carriers, and disease-associated glycolipid biomarkers). By addressing these challenges, AI-driven methods hold the potential to revolutionize the field of glyco-lipid biomacromolecule design.

### Property prediction

AI techniques have been widely used to predict the physicochemical properties, biological activities, and interaction characteristics of glycans and lipids. With the predicted properties, scientists design and synthesize glycans/lipids molecules with specific functions, reducing the cost of experiments and increasing the success rate.

Current AI models approach glycan sequence analysis through different perspectives: feature lists (e.g., GlyNet<sup>245</sup>), biological languages (e.g., SweetTalk<sup>246</sup> and glyBERT<sup>247</sup>), or molecular graphs (e.g., SweetNet<sup>248</sup>). These models incorporate adaptations to address the branched, nonlinear structure of glycans, resulting in significant improvements in prediction accuracy. The scope of glycan property prediction encompasses multiple aspects, including glycan type identification, classification, immunogenic potential assessment, and determination of associations with bacterial pathogenicity.<sup>246,248</sup> Through these trained models, researchers can predict properties of novel glycans and identify crucial motifs, such as the human-like glycan patterns that pathogenic *E. coli* strains use for molecular mimicry. Furthermore, researchers have implemented machine learning to examine the relationship between clinical characteristics and glycans in cancer patients.<sup>368</sup> The technology has also enabled the prediction of  $\alpha$ -fetoprotein (AFP)-negative hepatocellular carcinoma through glycan fingerprint analysis.<sup>369</sup>

AI techniques extend beyond glycan analysis to contribute significantly to lipid property prediction. In a groundbreaking study, Han et al.<sup>251</sup> combined machine learning with molecular modeling to predict crucial liposome formulation parameters. These parameters include size, polydispersity index (PDI), zeta potential, and encapsulation efficiency. Their research revealed that drug molecules with specific characteristics - optimal solubility (logS), molecular complexity, and XLogP3 values - serve as ideal candidates for highly efficient liposome preparation. In the field of metabolic lipid synthesis, AI has provided valuable insights into regulatory mechanisms. At the University of New South Wales, researchers utilized AI to map key interactions and regulatory pathways,<sup>252</sup> leading to the discovery of potential biomarkers and drug targets associated with metabolic diseases. The continuous advancement of AI technology

promises to further expand its applications in lipid metabolism research, with significant implications for precision medicine and understanding disease mechanism.

### Structure analysis and design

Accurate structural information helps understand the interactions between glycosphingolipids and other biomolecules (such as proteins and nucleic acids), revealing their specific roles in cellular signal transduction, immune response, and cell recognition. Understanding the architecture of glycosphingolipids can yield valuable scientific knowledge for the design of drugs, the diagnosis and treatment of diseases, and the formulation of vaccines, thus propelling advances in biomedical and biotechnological fields.

AI methods have evolved to provide sophisticated approaches analyzing and predicting the structures of glycan and lipid molecules. DeepGlyco<sup>249</sup> addresses the complexity of non-linear glycan structures by predicting fragment spectra of intact glycopeptides. This model integrates tree-structured LSTMs for glycan modeling with graph neural networks to capture fragmentation pathways, thereby enhancing explainability and improving the discrimination of glycan structural isomers. CandyCrunch<sup>250</sup> employs a two-step approach: initially training a dilated residual neural network using annotated MS/MS spectra, followed by glycan structure prediction from raw liquid chromatography-MS/MS data. For lipid analysis, LipidCCS<sup>253</sup> implements support vector regression (SVR) to achieve accurate predictions of collision cross-section (CCS) values. The method optimizes molecular descriptors to capture subtle structural variations among lipids, thus facilitating their identification. Beyond these specialized tools, AlphaFold's framework shows potential for expansion to predict both glycan and lipid structures, offering new possibilities for understanding their functions and molecular interactions.

AI-driven optimization of glycans and lipids has enabled significant advances across multiple domains. Key applications include the *de novo* design of ionizable lipids for mRNA delivery systems, optimized liposome drug carriers with machine learning-predicted formulation parameters such as size, polydispersity, and encapsulation efficiency,<sup>251,370</sup> and the identification of disease-associated glycolipid biomarkers through integrated analysis of metabolic gene signatures.<sup>371</sup> These AI-driven approaches enable the creation of highly targeted and efficient compounds that address specific clinical needs, advancing innovations in disease treatment and prevention strategies.

While researchers have primarily relied on traditional chemical and enzymatic synthesis for designing lipids, recent studies<sup>254</sup> have extended the application of AI from structural analysis to *de novo* generative design. A key example is the development of an adapted Synthesis-DAGs model,<sup>254</sup> a deep generative framework specifically for creating novel ionizable lipids, which are critical components of mRNA delivery systems. A core innovation of this model is its ability to address synthesizability by co-generating a molecule's structure along with its corresponding synthesis pathway, represented as a directed acyclic graph (DAG). By training on a custom-built dataset of over 70,000 synthesis paths derived from accessible building blocks and integrating an advanced reaction predictor, the model achieved an 83.4% success rate in generating valid ionizable lipids. Furthermore, it was demonstrated that the model can be iteratively optimized toward specific objectives, such as enhancing mRNA transfection efficiency. This work represents a critical step toward automated, goal-oriented lipid molecular design.

### Summary

The prediction of glycan and lipid molecular properties and their structural design presents significant technical challenges that continue to engage researchers. Despite the development of sophisticated algorithms, this field requires further innovation and systematic exploration. The structural design of these biomolecules is particularly crucial, given their fundamental role in optimizing human health outcomes. Current progress faces limitations due to the scarcity of high-quality experimental data, but this has also spurred innovative strategies, such as constructing domain-specific synthesis databases to train specialized generative models for lipid design, thus turning a key limitation into an opportunity for advancement. As comprehensive databases and resources become increasingly available, the structural design of glycans and lipids is emerging as a critical research frontier, promising transformative

breakthroughs in medicine and biotechnology.

## BIOMATERIALS

Biomaterials are materials designed to interact with biological systems for the replacement, repair, or regeneration of tissues and organs. They play a crucial role in advancing biomedical engineering, as they are foundational in applications such as medical devices, tissue engineering, and drug delivery systems.<sup>372–373</sup> Biomaterials are broadly classified into three main categories: polymers, metals, and ceramics. Among these, metal-based biomaterials hold the largest commercial market share, while ceramics follow closely behind. However, polymer-based materials represent the fastest-growing sector due to their versatility and adaptability. To meet clinical needs, biomaterials must exhibit excellent biocompatibility, mechanical properties, and degradability.<sup>374</sup> AI technology offers a dual approach to biomaterial development: predictive modeling of material properties (e.g., swelling behavior, gelation time, biocompatibility, and Young's modulus) and structural design optimization across polymeric, metallic, and ceramic biomaterials (Figure 4C).<sup>375</sup>

### AI-assisted prediction of biomaterial properties

AI techniques have been extensively applied to predict the properties of biomaterials, enabling efficient and accurate prediction of critical properties such as mechanical performance, biocompatibility, and degradation behavior. Polymeric biomaterials, particularly hydrogels, have received the most attention. Soleimani et al. used the artificial neural network-salp swarm algorithm (ANN-SSA) to analyze swelling characteristics,<sup>255</sup> while Islamkulov et al. developed an ANN-based tool for predicting swelling behavior of cross-linked hydrogels.<sup>261</sup> Boztepe et al. compared ANN, LS-SVM, and SVR models for simulating swelling performance of pH- and temperature-responsive hydrogels, demonstrating that ANN achieved the best predictive performance.<sup>262</sup> Karaoglu et al. used ANN to predict stiffness and gelation time of GelMA hydrogels,<sup>256</sup> Wang et al. employed back-propagation neural networks to predict swelling states of temperature-responsive hydrogels,<sup>257–258</sup> Khalvandi et al. used deep neural networks to predict compressive stress-strain responses,<sup>259</sup> and Zheng et al. applied hierarchical convolutional LSTM to predict fracture behavior.<sup>260</sup> Beyond hydrogels, Epa et al. linked polymer properties to biological effects such as cell adhesion,<sup>268</sup> Kwaria et al. predicted protein adsorption on self-assembled monolayers,<sup>269</sup> and Rostam et al. modeled immune-guiding polymers to predict foreign body responses.<sup>270</sup> In the domain of metallic biomaterials, Wu et al. developed ANN models to identify low-modulus titanium alloys suitable for bone implants,<sup>271</sup> and Chang et al. predicted compositions of high-entropy alloys to optimize hardness.<sup>272</sup>

### AI-driven structural design of biomaterials

Various AI methods have been applied to design and optimize biomaterials. Hydrogel design has received the most attention among polymeric biomaterials. Chen et al. proposed an AI-assisted high-throughput screening system (AI-HTPCSS) that integrates a programmable pneumatic extrusion bioprinter with AI-assisted image analysis for efficient screening of printing conditions.<sup>263</sup> Liu et al. developed a deep learning-based approach using GANs for designing superfiber-reinforced hydrogel composite structures optimized for specific mechanical performance.<sup>264</sup> Umar et al. combined DFT calculations with machine learning to design sodium alginate-based hydrogels for pollutant adsorption.<sup>265</sup> Qiu et al. integrated BPNN and multi-population genetic algorithms to rapidly design hydrogel-based metamaterials with targeted negative hydration expansion effects.<sup>266</sup> Cadamuro et al.<sup>267</sup> developed a neural network model for designing extracellular matrix-mimicking hydrogels with custom rheological properties, accurately predicting formulations based on target storage and loss modulus. Machine learning has also been extended to metallic and ceramic biomaterials. Xu et al.<sup>376</sup> combined physical metallurgical models with XGBoost algorithm to design  $\beta$ -type medical titanium alloys with low elastic modulus in the Ti-Mo-Nb-Zr-Sn system, achieving experimental elastic moduli of 37–43 GPa that closely match natural bone. Horikawa et al.<sup>377</sup> utilized machine learning to construct models that estimate the bone-forming ability of porous hydroxyapatite ceramics from fabrication conditions, enabling inverse analysis-based material design with experimentally validated biological responses.

## Summary

AI has been widely adopted in biomaterial research, though most applications to date focus on property prediction rather than direct *de novo* design. Existing AI methods for biomaterial design remain relatively conventional, often relying on standard optimization frameworks rather than cutting-edge generative AI techniques.

The application of GenAI in biomaterials remains significantly underdeveloped compared to other molecular design domains, stemming from fundamental challenges including the complexity of multi-scale properties, the lack of large standardized datasets, and the diversity of material representations across polymers, metals, and ceramics. To unlock the full potential of GenAI for biomaterial design, future efforts should focus on creating standardized high-quality datasets and improving AI model interpretability, paving the way for breakthroughs in biomedical engineering.

## FUTURE DIRECTIONS

AI-driven molecule design has made significant strides, particularly in the generation of protein structures and small molecules. Advances in GenAI models, such as GANs and deep learning architectures, have enabled the exploration of vast chemical spaces and the design of molecules with specific properties. These models have shown promise in tasks like drug discovery, protein engineering, and the generation of novel biomolecules. However, despite these advancements, the field still faces substantial challenges from the perspectives of data, AI models and evaluation.

From a data-centric perspective, lack of sufficient high-quality data would constrain the models' learning capacity and limit the diversity of generated molecules. Proprietary data from pharmaceutical companies, which includes small molecule structures and protein interaction data, is often confidential, restricting the learning scope of AI models. Public datasets, while available, suffer from inconsistent annotations and limited structural diversity, which can lead to biased or incomplete training. These limitations hinder the model's ability to learn general chemical rules and generate diverse, high-quality molecules.

From a model-centric perspective, several issues persist. Training stability and efficiency remain significant concerns, with models like GANs being prone to mode collapse, which reduces the diversity of generated molecules. Deep generative models also demand considerable computational resources, making them less accessible. Furthermore, ensuring the chemical validity and synthetic feasibility of generated molecules is a critical challenge, as models may produce compounds that violate fundamental chemical rules or are impractical to synthesize. The interpretability of these models are also limited, making it difficult to fine-tune specific molecular attributes or understand the reasoning behind generated outputs.

Evaluating the quality and utility of AI-generated molecules presents its own set of challenges. There is a lack of standardized metrics for assessing the novelty, quality, and practical applicability of generated molecules, making it difficult to compare different models or approaches. Experimental validation, while essential, is costly and time-consuming, slowing down the iterative improvement of models. Moreover, ensuring the trustworthiness and reliability of generated molecules is critical, as models may produce compounds that are chemically invalid, synthetically infeasible, or even unsafe.

We discussed the challenges for AI-driven molecule design from the above three perspectives. Next, we propose future research directions to address these challenges and advance the field further.

### Data-efficient AI for molecule design

The development of data-efficient AI models is crucial for molecule design, as high-quality molecular datasets are often limited, expensive to obtain, or proprietary. Current AI models rely heavily on large datasets, but real-world molecular discovery often involves small or noisy datasets. To address this limitation, data-efficient AI approaches,<sup>378</sup> such as few-shot learning, self-supervised learning, and transfer learning, can be leveraged to enhance the model's ability to generalize from limited data.

One promising direction is the application of self-supervised learning techniques that extract meaningful representations from unlabeled molecular data. Methods such as masked molecular modeling, contrastive learning, knowledge distillation, and graph-based pretraining allow models to learn



underlying molecular structures without extensive labeled datasets. In a recent study,<sup>379</sup> an end-to-end AI-guided few-shot inverse design framework leverages graph grammar knowledge distillation to explore novel polymer structures with desired properties. By distilling and recombining graph grammar fragments from limited polymer data, the framework efficiently restricts the high-dimensional polymer space, enabling the identification of 83 antimicrobial candidates from just 86 initial polymers. Additionally, integrating domain knowledge—such as chemical heuristics and bioactivity patterns—into generative models can further improve data efficiency. Finally, few-shot learning frameworks can be designed to adapt to novel molecular tasks using only a small number of labeled samples, making AI models more practical to drug discovery, material design, and other real-world applications.

Beyond traditional self-supervised learning approaches, recent advances in molecular representation learning offer concrete pathways for data efficiency. Masked molecular modeling techniques, such as masked atom prediction where models learn to predict randomly masked atoms in molecular graphs, can extract meaningful representations from unlabeled data. Multi-scale diffusion approaches like HGLDM<sup>222</sup> demonstrate how hierarchical diffusion processes can denoise simultaneously across multiple levels of molecular abstraction (atoms, functional groups, and complete molecules), enabling efficient learning from limited data by leveraging structural hierarchies. Additionally, cross-dimensional information transfer, as exemplified by the SyCo model,<sup>223</sup> provides a concrete framework where coarse 2D structural sketches serve as priors for 3D geometry learning, which then guides higher-fidelity 2D molecular graph generation—effectively amplifying the utility of limited training data through multi-modal learning.

### Foundation models for molecule design

Foundation models have demonstrated success in natural language processing and computer vision, and their extension to molecule design holds great promise. A foundation model for molecular generation and prediction can be trained on diverse chemical and biological datasets, enabling generalization across multiple downstream tasks such as drug discovery, material synthesis, protein engineering, etc.

These models can benefit from multi-modal learning, where molecular data from various sources—such as SMILES strings, molecular graphs, 3D structures, and experimental bioassays—are integrated to create a more comprehensive representation of chemical space. Transformer-based architectures, large-scale pretraining, and fine-tuning strategies will be essential for building robust foundation models that can generate novel molecules, predict their properties, and optimize them for specific applications. For example, ChemFM,<sup>380</sup> a large foundation model specifically developed for chemicals, can significantly enhance the properties of generated molecules in conditional molecular generation tasks. In addition, few-shot fine-tuning the foundation models<sup>381–383</sup> for downstream prediction and generation tasks is also well aligned with our above data-efficient AI for molecule generation. Lastly, RL-based optimization loops can be integrated to refine the generated molecules based on synthetic feasibility, bioactivity, and safety constraints.

### Physics/Chemistry-informed AI for molecule design

Incorporating physical and chemical principles into AI-driven molecule design can enhance model reliability and interpretability. Current deep learning methods often lack explicit knowledge of molecular physics, leading to the generation of molecules that may be chemically invalid or synthetically infeasible. To mitigate these issues, hybrid AI approaches that integrate physics-based simulations, quantum mechanics, and MD with deep learning can provide more accurate predictions.

Several physics-informed architectures have emerged for molecular design. Co-diffusion frameworks, such as MolDiff,<sup>225</sup> explicitly model the interdependencies between atoms and chemical bonds during the generation process, ensuring structural consistency that traditional atom-only DMs often violate. Multi-scale interaction weighting, implemented in MD3MD,<sup>226</sup> decomposes molecules into multi-scale graphs and assigns different weights to inter-atomic interactions at various scales, effectively incorporating physical constraints into the diffusion process. Specialized noising schemes for multimodal features, as demonstrated by EDM,<sup>99</sup> provide equivariant DMs that handle both discrete atomic types and continuous coordinates

while adjusting input feature scales for enhanced training stability. Additionally, retrosynthesis-guided generation, exemplified by TANGO,<sup>229</sup> integrates synthesis prediction tools directly into the reward function, fundamentally constraining generation to synthetically feasible molecules by limiting building blocks to predefined sets.

### Ethics in AI-driven molecule design

As AI-driven molecule design progresses, ethical considerations must be prioritized to ensure responsible and safe applications. The ability to generate novel molecules, including bioactive compounds and materials with unknown effects, raises concerns about misuse, such as the design of harmful chemicals or bioweapons. To address these risks, AI models must incorporate safeguards, such as controlled access to molecular generation tools, ethical review processes, and transparency in model outputs.

Bias and fairness in molecular datasets also pose challenges, as AI models trained on biased data may overlook certain molecular properties or fail to generalize across diverse chemical spaces. Ensuring dataset diversity and implementing bias-mitigation techniques will be critical to developing equitable AI-driven molecular design systems. Furthermore, regulatory frameworks must be established to monitor AI-generated molecules, particularly in pharmaceutical applications, to prevent unintended consequences and ensure compliance with safety standards.

### CONCLUSIONS

AI-driven molecule design has revolutionized the discovery and development of novel molecules across various scientific fields, including drug discovery, material science, and synthetic biology. This survey has explored the application of GenAI techniques within and beyond the central dogma, covering the design of DNAs, RNAs, proteins, small molecules, lipids, glycans, and biomaterials. Despite significant advancements, several challenges remain, including data scarcity, model interpretability, and experimental validation. Addressing these issues requires interdisciplinary collaboration among AI researchers, chemists, and biologists to refine existing methods and develop new frameworks. Future research should emphasize improving data efficiency, leveraging foundation models, incorporating physics-informed AI, and establishing ethical safeguards to ensure responsible AI-driven molecular innovation. In summary, AI has the potential to fundamentally reshape molecule design by accelerating discovery processes, improving accuracy, and uncovering novel molecular structures that were previously inaccessible. With continued advancements, AI-driven approaches would play a pivotal role in pushing the boundaries of molecular science and transforming the future of biotechnology and therapeutic development.

### REFERENCES

- Romanelli V., Cerchia C. and Lavecchia A. (2024). Unlocking the potential of generative artificial intelligence in drug discovery. Lyu Z. (ed.). *Applications of Generative AI* (Springer International Publishing), pp: 37–63. DOI:10.1007/978-3-031-46238-2\_3
- Tang X., Dai H., Knight E., et al. (2024). A survey of generative AI for *de novo* drug design: New frontiers in molecule and protein generation. *Brief. in Bioinform.* **25**:bbae338. DOI:10.1093/bib/bbae338
- Han S. and Wu J. (2024). Artificial intelligence (AI) meets biomaterials and biomedicine. *Smart Mater. in Med.* **5**:251–255. DOI:10.1016/j.smaim.2024.03.001
- Goshisht M.K. (2024). Machine learning and deep learning in synthetic biology: Key architectures, applications, and challenges. *ACS Omega* **9**:9921–9945. DOI:10.1021/acsomega.3c05913
- Shi Y., Zhen X., Zhang Y., et al. (2024). Chemically modified platforms for better RNA therapeutics. *Chem. Rev.* **124**:929–1033. DOI:10.1021/acs.chemrev.3c00611
- Yi Y., An H.W. and Wang H. (2024). Intelligent biomaterialomics: Molecular design, manufacturing, and biomedical applications. *Adv. Mater.* **36**:2305099. DOI:10.1002/adma.202305099
- Yang N., Kong B., Zhu Z., et al. (2024). Recent advances in targeted protein degraders as potential therapeutic agents. *Mol. Diversity* **28**:309–333. DOI:10.1007/s11030-023-10606-w
- Zhou Y., Qian X., Li J., et al. (2024). The role of tumor-associated macrophages in lung cancer: From mechanism to small molecule therapy. *Biomedicine & Pharmacother.* **170**:116014. DOI:10.1016/j.biopha.2023.116014
- Somasundaram N., Vigneshkumar M., Pawar S.R., et al. (2024). AI-driven material design for tissue engineering: a comprehensive approach integrating generative adversarial networks and high-throughput experimentation. *The Sci. Temper*

- 15:1576–1580. DOI:10.58414/scientifictemper.2024.15.1.02
10. Liu M., Li C., Chen R., et al. (2024). Geometric deep learning for drug discovery. *Expert Syst. with Appl.* **240**:122498. DOI:10.1016/j.eswa.2023.122498
11. Bai Q., Xu T., Huang J., et al. (2024). Geometric deep learning methods and applications in 3D structure-based drug design. *Drug Discov. Today* **29**:104024. DOI:10.1016/j.drudis.2024.104024
12. Zheng S., He J., Liu C., et al. (2024). Predicting equilibrium distributions for molecular systems with deep learning. *Nat. Mach. Intell.* **6**:558–567. DOI:10.1038/s42256-024-00837-3
13. Guan X., Tang Q.Y., Ren W., et al. (2024). Predicting protein conformational motions using energetic frustration analysis and AlphaFold2. *Proc. Natl. Acad. Sci.* **121**:e2410662121. DOI:10.1073/pnas.2410662121
14. Notin P., Rollins N., Gal Y., et al. (2024). Machine learning for functional protein design. *Nat. Biotechnol.* **42**:216–228. DOI:10.1038/s41587-024-02127-0
15. Kortemme T. (2024). *De novo* protein design—From new structures to programmable functions. *Cell* **187**:526–544. DOI:10.1016/j.cell.2023.12.028
16. Zhong W., Kim H. and Kim W.Y. (2024). 3D molecular generative framework for interaction-guided drug design. *Nat. Commun.* **15**. DOI:10.1038/s41467-024-47011-2
17. Dauparas J., Anishchenko I., Bennett N., et al. (2022). Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**:49–56. DOI:10.1101/2022.06.03.494563
18. Moret M., Friedrich L., Grisoni F., et al. (2020). Generative molecular design in low data regimes. *Nat. Mach. Intell.* **2**:171–180. DOI:10.1038/s42256-020-0160-y
19. Anand N., Eguchi R., Mathews I.I., et al. (2022). Protein sequence design with a learned potential. *Nat. Commun.* **13**:746. DOI:10.1038/s41467-022-28313-9
20. Syrlybaeva R. and Strauch E.M. (2023). Deep learning of protein sequence design of protein–protein interactions. *Bioinformatics* **39**:btac733. DOI:10.1093/bioinformatics/btac733
21. Meyers J., Fabian B. and Brown N. (2021). *De novo* molecular design and generative models. *Drug Discov. Today* **26**:2707–2715. DOI:10.1016/j.drudis.2021.05.019
22. Luo S., Guan J., Ma J., et al. (2021). A 3D generative model for structure-based drug design. In *Adv. Neural Inf. Process. Syst.* **34**: 6229–6239. [https://proceedings.neurips.cc/paper\\_files/paper/2021/file/314450613369e0ee72d0da7f6fee773c-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2021/file/314450613369e0ee72d0da7f6fee773c-Paper.pdf)
23. Jin W., Wohlwend J., Barzilay R., et al. (2021). Iterative refinement graph neural network for antibody sequence-structure co-design. *arXiv abs/2110.04624*. DOI:10.48550/arXiv.2110.04624
24. Wu F. and Li S.Z. (2023). A hierarchical training paradigm for antibody structure-sequence co-design. *Adv. in Neural Inf. Processing Syst.* **36**:31140–31157. DOI:10.1101/2022.11.14.516404
25. Elton D.C., Boukouvalas Z., Fuge M.D., et al. (2019). Deep learning for molecular design—A review of the state of the art. *Mol. Syst. Des. & Eng.* **4**:828–849. DOI:10.1039/c9me00039a
26. Shen J. and Nicolaou C.A. (2019). Molecular property prediction: Recent trends in the era of artificial intelligence. *Drug Discov. Today: Technol.* **32**:29–36. DOI:10.1016/j.ddtec.2020.05.001
27. Zhang J., Lei Y.K., Zhang Z., et al. (2020). A perspective on deep learning for molecular modeling and simulations. *The J. of Phys. Chem. A* **124**:6745–6763. DOI:10.1021/acs.jpcc.0c04473
28. Preuer K., Klambauer G., Rippmann F., et al. (2019). Interpretable deep learning in drug discovery. Samek W., Montavon G., Vedaldi A., et al. (eds.). *Explainable AI: Interpreting, explaining and visualizing deep learning* (Springer), pp: 331–345. DOI:10.1007/978-3-030-28954-6\_18
29. Li Y., Zhang L. and Liu Z. (2018). Multi-objective *de novo* drug design with conditional graph generative model. *J. of Cheminformatics* **10**:1–24. DOI:10.1186/s13321-018-0287-6
30. Dash T., Srinivasan A., Vig L., et al. (2022). Using domain-knowledge to assist lead discovery in early-stage drug design. Katouris N. and Artikis A. (eds.). *International Conference on Inductive Logic Programming* (Springer International Publishing), pp: 78–94. DOI:10.1007/978-3-030-97454-1\_6
31. Chu A.E., Lu T. and Huang P.S. (2024). Sparks of function by *de novo* protein design. *Nat. Biotechnol.* **42**:203–215. DOI:10.1038/s41587-024-02133-2
32. Li G., Chen C., Li Y., et al. (2024). DNA-origami-based precise molecule assembly and their biological applications. *Nano Lett.* **24**:11335–11348. DOI:10.1021/acs.nanolett.4c03297
33. Cao X., Zhang Y., Ding Y., et al. (2024). Identification of RNA structures and their roles in RNA functions. *Nat. Rev. Mol. Cell Biol.* **25**:784–801. DOI:10.1038/s41580-024-00748-6
34. Zhou J. and Huang M. (2024). Navigating the landscape of enzyme design: From molecular simulations to machine learning. *Chem. Soc. Rev.* **53**:8202–8239. DOI:10.1039/d4cs00196f
35. Frappier V. and Keating A.E. (2021). Data-driven computational protein design. *Curr. Opin. in Struct. Biol.* **69**:63–69. DOI:10.1016/j.sbi.2021.03.009
36. Martinelli D.D. (2022). Generative machine learning for *de novo* drug discovery: A systematic review. *Computers in Biol. and Med.* **145**:105403. DOI:10.1016/j.combiomed.2022.105403
37. Cheng Y., Gong Y., Liu Y., et al. (2021). Molecular design in drug discovery: A comprehensive review of deep generative models. *Brief. in Bioinform.* **22**:bbab344. DOI:10.1093/bib/bbab344
38. Zeng X., Wang F., Luo Y., et al. (2022). Deep generative molecular design reshapes drug discovery. *Cell Rep. Med.* **3**:100794. DOI:10.1016/j.xcrm.2022.100794
39. Zhang S., Fan R., Liu Y., et al. (2023). Applications of transformer-based language models in bioinformatics: A survey. *Bioinform. Adv.* **3**:vbad001. DOI:10.1093/bioadv/vbad001
40. Guo Z., Liu J., Wang Y., et al. (2024). Diffusion models in bioinformatics and computational biology. *Nat. Rev. Bioeng.* **2**:136–154. DOI:10.1038/s44222-023-00114-9
41. Lan L., You L., Zhang Z., et al. (2020). Generative adversarial networks and its applications in biomedical informatics. *Front. in Public Health* **8**:164. DOI:10.3389/fpubh.2020.00164
42. Wei R. and Mahmood A. (2020). Recent advances in variational autoencoders with representation learning for biomedical informatics: A survey. *IEEE Access* **9**:4939–4956. DOI:10.1109/access.2020.3048309
43. Sankaran K. and Holmes S.P. (2023). Generative models: An interdisciplinary perspective. *Annu. Rev. of Stat. and Its Appl.* **10**:325–352. DOI:10.1146/annurev-statistics-033121-110134
44. Rombach R., Blattmann A., Lorenz D., et al. (2022). High-resolution image synthesis with latent diffusion models. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 10684–10695. DOI:10.1109/cvpr52688.2022.01042
45. Radford A., Wu J., Child R., et al. (2019). Language models are unsupervised multitask learners. OpenAI Technical Report. URL <https://openai.com/index/better-language-models>
46. Liao R., Li Y., Song Y., et al. (2019). Efficient graph generation with graph recurrent attention networks. In *Adv. Neural Inf. Process. Syst.* **32**. [https://proceedings.neurips.cc/paper\\_files/paper/2019/file/d0921d442ee91b896ad95059d13df618-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2019/file/d0921d442ee91b896ad95059d13df618-Paper.pdf)
47. Huang R., Cui C., Chen F., et al. (2022). Singgan: Generative adversarial network for high-fidelity singing voice generation. In *Proceedings of the 30th ACM International Conference on Multimedia* 2525–2535. DOI:10.1145/3503161.3547854
48. Wang D., Xiao M., Wu M., et al. (2023). Reinforcement-enhanced autoregressive feature transformation: Gradient-steered search in continuous space for postfix expressions. In *Adv. Neural Inf. Process. Syst.* **36**: 43563–43578. [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/8797d13e5998acfab387d4bf0a5b9b00-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/8797d13e5998acfab387d4bf0a5b9b00-Paper-Conference.pdf)
49. Wang H., Zhang Z., Wang T., et al. (2023). PrivTrace: Differentially private trajectory synthesis by adaptive Markov models. In *32nd USENIX Security Symposium (USENIX Security 23)* 1649–1666. <https://www.usenix.org/conference/usenixsecurity23/presentation/wang-haiming>
50. Bourou S., El Saer A., Velivassaki T.H., et al. (2021). A review of tabular data synthesis using GANs on an IDS dataset. *Inf.* **12**:375. DOI:10.3390/info12090375
51. LeCun Y., Chopra S., Hadsell R., et al. (2006). A tutorial on energy-based learning. Predicting Structured Data. <http://yann.lecun.com/exdb/publis/pdf/lecun-06.pdf>
52. Fischer A. and Igel C. (2012). An introduction to restricted Boltzmann machines. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications* 14–36. DOI:10.1007/978-3-642-33275-3\_2
53. Hinton G.E. (2009). Deep belief networks. *Scholarpedia* **4**:5947. DOI:10.4249/scholarpedia.5947
54. Ngiam J., Chen Z., Koh P.W., et al. (2011). Learning deep energy models. In *International Conference on Machine Learning* 1105–1112. [https://icml.cc/2011/papers/557\\_icmlpaper.pdf](https://icml.cc/2011/papers/557_icmlpaper.pdf)
55. Zhai S., Cheng Y., Lu W., et al. (2016). Deep structured energy based models for anomaly detection. In *International Conference on Machine Learning* **48**: 1100–1109. <https://proceedings.mlr.press/v48/zhai16.html>
56. Du Y. and Mordatch I. (2019). Implicit generation and modeling with energy based models. In *Advances in Neural Information Processing Systems* **32**. [https://proceedings.neurips.cc/paper\\_files/paper/2019/file/378a063b8fdb1db941e34f4bde584c7d-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2019/file/378a063b8fdb1db941e34f4bde584c7d-Paper.pdf)
57. Liu M., Yan K., Oztekin B., et al. (2021). GraphEBM: Molecular graph generation with energy-based models. In *Energy Based Models Workshop - ICLR 2021* URL <https://openreview.net/forum?id=Gc51PtLzYw>
58. Kong D., Pang B., Han T., et al. (2023). Molecule design by latent space energy-based modeling and gradual distribution shifting. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence* **216**: 1109–1120. <https://proceedings.mlr.press/v216/kong23a.html>
59. Bengio Y., Courville A. and Vincent P. (2013). Representation learning: A review and new perspectives. *IEEE Trans. on Pattern Anal. and Mach. Intell.* **35**:1798–1828. DOI:10.1109/tpami.2013.50
60. Hinton G.E. and Salakhutdinov R.R. (2006). Reducing the dimensionality of data with neural networks. *Science* **313**:504–507. DOI:10.1126/Science.1127647
61. Kingma D.P. (2013). Auto-encoding variational bayes. *arXiv abs/1312.6114*. DOI:10.48550/arXiv.1312.6114
62. Kingma D.P., Rezende D.J., Mohamed S., et al. (2014). Semi-supervised learning with deep generative models. *Adv. in Neural Inf. Processing Syst.* **27**. [https://proceedings.neurips.cc/paper\\_files/paper/2014/file/6d42b1217a6996997ea5d5a8398c1f944-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2014/file/6d42b1217a6996997ea5d5a8398c1f944-Paper.pdf)
63. Sønderby C.K., Raiko T., Maaløe L., et al. (2016). Ladder variational autoencoders. In

- Adv. Neural Inf. Process. Syst. **29**. [https://proceedings.neurips.cc/paper\\_files/paper/2016/file/6ae07dcb33ec3b7c814df797cbda0f87-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2016/file/6ae07dcb33ec3b7c814df797cbda0f87-Paper.pdf)
64. Higgins I., Matthey L., Pal A., et al. (2017).  $\beta$ -VAE: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations* <https://openreview.net/forum?id=Sy2fzU9gl>
  65. Hong S.H., Ryu S., Lim J., et al. (2019). Molecular generative model based on an adversarially regularized autoencoder. *J. of Chem. Inf. and Model.* **60**:29–36. DOI:10.1021/acs.jcim.9b00694
  66. Nash C. and Williams C.K. (2017). The shape variational autoencoder: A deep generative model of part-segmented 3D objects. *Comput. Graph. Forum* **36**:1–12. DOI:10.1111/cgf.13240
  67. Gómez-Bombarelli R., Wei J.N., Duvenaud D., et al. (2018). Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Sci.* **4**:268–276. DOI:10.1021/acscentsci.7b00572
  68. Kusner M.J., Paige B. and Hernández-Lobato J.M. (2017). Grammar variational autoencoder. In *International Conference on Machine Learning* **70**: 1945–1954. <https://proceedings.mlr.press/v70/kusner17a.html>
  69. Dinh L., Sohl-Dickstein J. and Bengio S. (2016). Density estimation using Real NVP. *arXiv abs/1605.08803*. DOI:10.48550/arXiv.1605.08803
  70. Kingma D.P. and Dhariwal P. (2018). Glow: Generative flow with invertible 1x1 convolutions. In *Adv. Neural Inf. Process. Syst.* **31**. [https://proceedings.neurips.cc/paper\\_files/paper/2018/file/d139db6a236200b21cc7f752979132d0-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2018/file/d139db6a236200b21cc7f752979132d0-Paper.pdf)
  71. Grathwohl W., Chen R.T., Bettencourt J., et al. (2018). FFLORD: Free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations* <https://openreview.net/forum?id=rJxgknCck7>
  72. Madhawa K., Ishiguro K., Nakago K., et al. (2019). GraphNVP: An invertible flow model for generating molecular graphs. *arXiv abs/1905.11600*. DOI:10.48550/arXiv.1905.11600
  73. Geffner T., Didi K., Zhang Z., et al. (2025). Proteina: Scaling flow-based protein structure generative models. In *The Thirteenth International Conference on Learning Representations* <https://openreview.net/forum?id=TVQLu34bdw>
  74. Creswell A., White T., Dumoulin V., et al. (2018). Generative adversarial networks: An overview. *IEEE Signal Processing Mag.* **35**:53–65. DOI:10.1109/msp.2017.2765202
  75. Goodfellow I., Pouget-Abadie J., Mirza M., et al. (2014). Generative adversarial nets. In *Adv. Neural Inf. Process. Syst.* **27**. [https://proceedings.neurips.cc/paper\\_files/paper/2014/file/f033ed80deb0234979a61f95710dbe25-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2014/file/f033ed80deb0234979a61f95710dbe25-Paper.pdf)
  76. Mirza M. and Osindero S. (2014). Conditional generative adversarial nets. *arXiv abs/1411.1784*. DOI:10.48550/arXiv.1411.1784
  77. Lai W.S., Huang J.B., Ahuja N., et al. (2017). Deep laplacian pyramid networks for fast and accurate super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 5835–5843. DOI:10.1109/cvpr.2017.618
  78. Lin J., Zhang R., Ganz F., et al. (2021). Anycost GANs for interactive image synthesis and editing. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 14986–14996. DOI:10.1109/cvpr46437.2021.01474
  79. Karras T., Laine S., Aittala M., et al. (2020). Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 8107–8116. DOI:10.1109/cvpr42600.2020.00813
  80. Cao N.D. and Kipf T. (2018). MolGAN: An implicit generative model for small molecular graphs. *arXiv abs/1805.11973*. DOI:10.48550/arXiv.1805.11973
  81. Repecka D., Jauniskis V., Karpus L., et al. (2021). Expanding functional protein sequence spaces using generative adversarial networks. *Nat. Mach. Intell.* **3**:324–333. DOI:10.1038/s42256-021-00310-5
  82. Xing Z., Feng Q., Chen H., et al. (2024). A survey on video diffusion models. *ACM Comput. Surv.* **57**. DOI:10.1145/3696415 DOI:10.1145/3696415
  83. Ho J., Jain A. and Abbeel P. (2020). Denoising diffusion probabilistic models. *Adv. in Neural Inf. Processing Syst.* **33**: 6840–6851. [https://proceedings.neurips.cc/paper\\_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf)
  84. Song Y., Sohl-Dickstein J., Kingma D.P., et al. (2021). Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations* <https://openreview.net/forum?id=PXTIG12RRHS>
  85. Oksendal B. (2013). Stochastic differential equations. Stochastic differential equations: An introduction with applications (Springer Science & Business Media), pp: 65–84. DOI:10.1007/978-3-642-14394-6
  86. Rombach R., Blattmann A., Lorenz D., et al. (2022). High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 10674–10685. DOI:10.1109/cvpr52688.2022.01042
  87. Esser P., Kulal S., Blattmann A., et al. (2024). Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning* **235**: 12606–12633. URL <https://proceedings.mlr.press/v235/esser24a.html>
  88. Niu C., Song Y., Song J., et al. (2020). Permutation invariant graph generation via score-based generative modeling. In *International Conference on Artificial Intelligence and Statistics* **108**: 4474–4484. <https://proceedings.mlr.press/v108/niu20a.html>
  89. Hoogeboom E., Satorras V.G., Vignac C., et al. (2022). Equivariant diffusion for molecule generation in 3D. In *Proceedings of the 39th International Conference on Machine Learning* **162**: 8867–8887. <https://proceedings.mlr.press/v162/hoogeboom22a.html>
  90. Zhao W.X., Zhou K., Li J., et al. (2023). A survey of large language models. *arXiv abs/2303.18223*. DOI:10.48550/arXiv.2303.18223
  91. Zhou C., Li Q., Li C., et al. (2024). A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *Int. J. of Mach. Learn. and Cybernetics* **16**:9851–9915. DOI:10.1007/s13042-024-02443-6
  92. Vaswani A., Shazeer N., Parmar N., et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* **30**. [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf)
  93. Radford A., Narasimhan K., Salimans T., et al. (2018). Improving language understanding by generative pre-training. OpenAI Technical Report. <https://openai.com/index/language-unsupervised/>
  94. Devlin J., Chang M.W., Lee K., et al. (2019). BERT: Pre-training of deep bidirectional Transformers for language understanding. *arXiv abs/1810.04805*. DOI:10.48550/arXiv.1810.04805
  95. Brown T., Mann B., Ryder N., et al. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems* **33**: 1877–1901. [https://proceedings.neurips.cc/paper\\_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf)
  96. Achiam J., Adler S., Agarwal S., et al. (2023). GPT-4 technical report. *arXiv abs/2303.08774*. DOI:10.48550/arXiv.2303.08774
  97. OpenAI (2024). GPT-4o. OpenAI Technical Report. <https://openai.com/index/hello-gpt-4o/>
  98. Yang Z., Dai Z., Yang Y., et al. (2019). XLNet: Generalized autoregressive pretraining for language understanding. In *Adv. Neural Inf. Process. Syst.* **32**. [https://proceedings.neurips.cc/paper\\_files/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf)
  99. Du Z., Qian Y., Liu X., et al. (2022). GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* **1**: 320–335. DOI:10.18653/v1/2022.acl-long.26
  100. Zhang S., Roller S., Goyal N., et al. (2022). OPT: Open pre-trained transformer language models. *arXiv abs/2205.01068*. DOI:10.48550/arXiv.2205.01068
  101. Touvron H., Lavril T., Izacard G., et al. (2023). LLaMA: Open and efficient foundation language models. *arXiv abs/2302.13971*. DOI:10.48550/arXiv.2302.13971
  102. Grattafiori A., Dubey A., Jauhri A., et al. (2024). The LLaMA 3 herd of models. *arXiv abs/2407.21783*. DOI:10.48550/arXiv.2407.21783
  103. Sarkar A., Kang Y., Somia N., et al. (2024). Designing DNA with tunable regulatory activity using score-entropy discrete diffusion. *bioRxiv* DOI:10.1101/2024.05.23.595630
  104. Wu W., Li Q., Li M., et al. (2025). GENERator: A long-context generative genomic foundation model. *arXiv abs/2502.07272*. DOI:10.48550/arXiv.2502.07272
  105. Iñiguez-Muñoz S., Llinàs-Arias P., Ensenyat-Mendez M., et al. (2024). Hidden secrets of the cancer genome: Unlocking the impact of non-coding mutations in gene regulatory elements. *Cell. and Mol. Life Sci.* **81**:274. DOI:10.1007/s00018-024-05314-z
  106. Wang Y., Wang H., Wei L., et al. (2020). Synthetic promoter design in Escherichia coli based on a deep generative network. *Nucleic Acids Res.* **48**:6403–6412. DOI:10.1093/nar/gkaa325
  107. Cheng Y., Zhang X., Liu F., et al. (2024). PromoterDiff: De novo design approach for Escherichia coli promoters based on a diffusion model. In *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)* 2381–2387. DOI:10.1109/cscwd61410.2024.10580553
  108. Wang H., Du Q., Wang Y., et al. (2024). GPro: Generative AI-empowered toolkit for promoter design. *Bioinformatics* **40**:btac123. DOI:10.1093/bioinformatics/btac123
  109. Yang G., Li J., Hu J., et al. (2024). Recognition of cyanobacteria promoters via Siamese network-based contrastive learning under novel non-promoter generation. *Brief. in Bioinform.* **25**:bbac193. DOI:10.1093/bib/bbac193
  110. Zhang P., Wang H., Xu H., et al. (2023). Deep flanking sequence engineering for efficient promoter design using DeepSEED. *Nat. Commun.* **14**:6309. DOI:10.1038/s41467-023-41899-y
  111. Seo E., Choi Y.N., Shin Y.R., et al. (2023). Design of synthetic promoters for cyanobacteria with generative deep-learning model. *Nucleic Acids Res.* **51**:7071–7082. DOI:10.1093/nar/gkad451
  112. Lin J., Wang X., Liu T., et al. (2024). Diffusion-based generative network for de novo synthetic promoter design. *ACS Synth. Biol.* **13**:1513–1522. DOI:10.1021/acssynbio.4c00041
  113. Xia Y., Du X., Liu B., et al. (2024). Species-specific design of artificial promoters by transfer-learning based generative deep-learning model. *Nucleic Acids Res.* **52**:6145–6157. DOI:10.1101/2023.12.27.573416
  114. Wang X., Xu K., Tan Y., et al. (2023). Deep learning-assisted design of novel promoters in Escherichia coli. *Adv. Genet.* **4**:2300184. DOI:10.1002/ggn2.202300184
  115. Almeida B.P., Reiter F., Pagani M., et al. (2022). DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers.



- Nat. Genet.* **54**:613–624. DOI:10.1038/s41588-022-01048-5
116. Li Z., Zhang Y., Peng B., et al. (2024). A novel interpretable deep learning-based computational framework designed synthetic enhancers with broad cross-species activity. *Nucleic Acids Res.* **52**:13447–13468. DOI:10.1093/nar/gkae912
117. Yin C., Castillo-Hair S., Byeon G.W., et al. (2025). Iterative deep learning design of human enhancers exploits condensed sequence grammar to achieve cell-type specificity. *Cell Syst.* **16**. DOI:10.1101/2024.06.14.599076 DOI:10.1101/2024.06.14.599076
118. Yang Z., Su B., Cao C., et al. (2025). Regulatory DNA sequence design with reinforcement learning. *arXiv abs/2503.07981*. DOI:10.48550/arXiv.2503.07981
119. Wang C., Uehara M., He Y., et al. (2024). Fine-tuning discrete diffusion models via reward optimization with applications to DNA and protein design. *arXiv abs/2410.13643*. DOI:10.48550/arXiv.2410.13643
120. Feng C.Q., Zhang Z.Y., Zhu X.J., et al. (2019). iTerm-PseKNC: A sequence-based tool for predicting bacterial transcriptional terminators. *Bioinformatics* **35**:1469–1477. DOI:10.1093/bioinformatics/bty827
121. Cui W., Lin Q., Hu R., et al. (2021). Data-driven and *in silico*-assisted design of broad host-range minimal intrinsic terminators adapted for bacteria. *ACS Synth. Biol.* **10**:1438–1450. DOI:10.1021/acssynbio.1c00050
122. Li J., Wu L.F., Liu K., et al. (2025). Intelligent design of Escherichia coli terminators by coupling prediction and generation models. *ACS Synth. Biol.* **14**:3744–3752. DOI:10.1021/acssynbio.5c00429
123. Zeng W., Chen S., Cui X., et al. (2021). SilencerDB: A comprehensive database of silencers. *Nucleic Acids Res.* **49**:D221–D228. DOI:10.1093/nar/gkaa839
124. Pan J.H. and Du P.F. (2024). SilenceREIN: Seeking silencers on anchors of chromatin loops by deep graph neural networks. *Brief. in Bioinform.* **25**:bbad494. DOI:10.1093/bib/bbad494
125. Kim G.B., Gao Y., Palsson B.O., et al. (2021). DeepTFactor: A deep learning-based tool for the prediction of transcription factors. *Proc. Natl. Acad. Sci.* **118**:e2021171118. DOI:10.1073/pnas.2021171118
126. Liu J., Dong C., Jiang G., et al. (2020). Transcription factor expression as a predictor of colon cancer prognosis: A machine learning practice. *BMC Med. Genomics* **13**:1–10. DOI:10.1186/s12920-020-00775-0
127. Zhang Y., Liu T., Hu X., et al. (2021). CellCall: Integrating paired ligand–receptor and transcription factor activities for cell–cell communication. *Nucleic Acids Res.* **49**:8520–8534. DOI:10.1093/nar/gkab638
128. Alerasool N., Leng H., Lin Z.Y., et al. (2022). Identification and functional characterization of transcriptional activators in human cells. *Mol. Cell* **82**:677–695. DOI:10.1016/j.molcel.2021.12.008
129. Joshi C.K., Jamasb A.R., Viñas R., et al. (2024). gRNAde: Geometric deep learning for 3D RNA inverse design. *bioRxiv* DOI:10.1101/2024.03.31.587283
130. Wong F., He D., Krishnan A., et al. (2024). Deep generative design of RNA aptamers using structural predictions. *Nat. Comput. Sci.* **4**:829–839. DOI:10.1038/s43588-024-00720-6
131. Li Y., Pan X., Shen H., et al. (2025). DRAG: Design RNAs as hierarchical graphs with reinforcement learning. *Brief. in Bioinform.* **26**:bbaf106. DOI:10.1093/bib/bbaf106
132. Runge F., Franke J., Fertmann D., et al. (2024). Partial RNA design. *Bioinformatics* **40**:i437–i445. DOI:10.1093/bioinformatics/btae222
133. Huang H., Lin Z., He D., et al. (2024). RiboDiffusion: Tertiary structure-based RNA inverse folding with generative diffusion models. *bioRxiv* DOI:10.1101/2024.04.18.590187
134. Zhang H., Liu H., Xu Y., et al. (2025). Deep generative models design mRNA sequences with enhanced translational capacity and stability. *Science* **390**:eadr8470. DOI:10.1126/Science.adr8470
135. Zhao Y., Oono K., Takizawa H., et al. (2024). GenerRNA: A generative pre-trained language model for *de novo* RNA design. *PLoS One* **19**:e0310814. DOI:10.1371/journal.pone.0310814
136. Shukueian Tabrizi S., Barazandeh S., Hashemi Aghdam H., et al. (2025). RNAtranslator: Modeling protein-conditional RNA design as sequence-to-sequence natural language translation. *PLoS Comput. Biol.* **21**:e1013541. DOI:10.1371/journal.pcbi.1013541
137. Klypa R., Bietti A. and Grudin S. (2025). BANg: Bidirectional anchored generation for conditional RNA design. *arXiv abs/2502.21274*. DOI:10.48550/arXiv.2502.21274
138. Patel S., Peng F.Z., Fraser K., et al. (2025). EvoFlow-RNA: Generating and representing non-coding RNA with a language model. *bioRxiv* DOI:10.1101/2025.02.25.639942
139. Gao L. and Lu Z.J. (2024). RNACG: A universal RNA sequence conditional generation model based on flow-matching. *arXiv abs/2407.19838*. DOI:10.48550/arXiv.2407.19838
140. Castillo-Hair S.M. and Seelig G. (2021). Machine learning for designing next-generation mRNA therapeutics. *Accounts of Chem. Res.* **55**:24–34. DOI:10.1021/acs.accounts.1c00621
141. Baek M., McHugh R., Anishchenko I., et al. (2024). Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA. *Nat. Methods* **21**:117–121. DOI:10.1101/2022.09.09.507333
142. Hie B.L., Shanker V.R., Xu D., et al. (2024). Efficient evolution of human antibodies from general protein language models. *Nat. Biotechnol.* **42**:275–283. DOI:10.1038/s41587-023-01763-2
143. Bennett N.R., Watson J.L., Ragotte R.J., et al. (2026). Atomically accurate *de novo* design of antibodies with RFdiffusion. *Nature* **649**:183–193. DOI:10.1038/s41586-025-09721-5
144. Parkinson J., Hard R. and Wang W. (2023). The RESP AI model accelerates the identification of tight-binding antibodies. *Nat. Commun.* **14**:454. DOI:10.1038/s41467-023-36028-8
145. Mason D.M., Friedensohn S., Weber C.R., et al. (2021). Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nat. Biomed. Eng.* **5**:600–612. DOI:10.1038/s41551-021-00699-9
146. Makowski E.K., Kinnunen P.C., Huang J., et al. (2022). Co-optimization of therapeutic antibody affinity and specificity using machine learning models that generalize to novel mutational space. *Nat. Commun.* **13**:3788. DOI:10.1038/s41467-022-31457-3
147. Shanehsazadeh A., Bachas S., McPartlon M., et al. (2023). Unlocking *de novo* antibody design with generative artificial intelligence. *bioRxiv* DOI:10.1101/2023.01.08.523187
148. Shanker V.R., Bruun T.U., Hie B.L., et al. (2024). Unsupervised evolution of protein and antibody complexes with a structure-informed language model. *Science* **385**:46–53. DOI:10.1126/Science.adk8946
149. Desautels T.A., Arrildt K.T., Zemla A.T., et al. (2024). Computationally restoring the potency of a clinical antibody against omicron. *Nature* **629**:878–885. DOI:10.1038/s41586-024-07385-1
150. He H., He B., Guan L., et al. (2024). *De novo* generation of SARS-CoV-2 antibody CDRH3 with a pre-trained generative large language model. *Nat. Commun.* **15**:6867. DOI:10.1038/s41467-024-50903-y
151. Eguchi R.R., Choe C.A. and Huang P.S. (2022). Ig-VAE: Generative modeling of protein structure by direct 3D coordinate generation. *PLoS Comput. Biol.* **18**:e1010271. DOI:10.1371/journal.pcbi.1010271
152. Li C., Sutherland D., Hammond S.A., et al. (2022). AMPLify: Attentive deep learning model for discovery of novel antimicrobial peptides effective against WHO priority pathogens. *BMC Genomics* **23**:77. DOI:10.1101/2020.06.16.155705
153. Lee H., Lee S., Lee I., et al. (2023). AMP-BERT: Prediction of antimicrobial peptide function based on a BERT model. *Protein Sci.* **32**:e4529. DOI:10.1002/pro.4529
154. Losin L. and Veltri D. (2021). Exploring target specificity of antimicrobial peptides through deep learning embeddings. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics* DOI:10.1145/3459930.3469506
155. Sharma R., Shrivastava S., Singh S.K., et al. (2023). Artificial intelligence-based model for predicting the minimum inhibitory concentration of antibacterial peptides against ESKAPEE pathogens. *IEEE J. of Biomed. and Health Inform.* **28**:1949–1958. DOI:10.1109/jbhi.2023.3271611
156. Sharma R., Shrivastava S., Singh S.K., et al. (2023). EnDL-HemoLyt: Ensemble deep learning-based tool for identifying therapeutic peptides with low hemolytic activity. *IEEE J. of Biomed. and Health Inform.* **28**:1896–1905. DOI:10.1109/jbhi.2023.3264941
157. Teimouri H., Medvedeva A. and Kolomeisky A.B. (2023). Bacteria-specific feature selection for enhanced antimicrobial peptide activity predictions using machine-learning methods. *J. of Chem. Inf. and Model.* **63**:1723–1733. DOI:10.1021/acs.jcim.2c01551
158. Vishnepolsky B., Grigolava M., Managadze G., et al. (2022). Comparative analysis of machine learning algorithms on the microbial strain-specific AMP prediction. *Brief. in Bioinform.* **23**:bbac233. DOI:10.1101/2022.01.28.478081
159. Salem M., Keshavarzi Arshadi A. and Yuan J.S. (2022). AMPDeep: Hemolytic activity prediction of antimicrobial peptides using transfer learning. *BMC Bioinform.* **23**:389. DOI:10.21203/rs.3.rs-1615895/v1
160. Ansari M. and White A.D. (2023). Serverless prediction of peptide properties with recurrent neural networks. *J. of Chem. Inf. and Model.* **63**:2546–2553. DOI:10.1101/2022.05.18.492545
161. Ansari M. and White A.D. (2024). Learning peptide properties with positive examples only. *Digital Discov.* **3**:977–986. DOI:10.1039/d3dd00218g
162. Ferrell J.B., Remington J.M., Van Oort C.M., et al. (2020). A generative approach toward precision antimicrobial peptide design. *bioRxiv* DOI:10.1101/2020.10.02.324087
163. Van Oort C.M., Ferrell J.B., Remington J.M., et al. (2021). AMPGAN v2: Machine learning-guided design of antimicrobial peptides. *J. Chem. Inf. Model.* **61**:2198–2207. DOI:10.1101/2020.11.18.388843
164. Buehler M.J. (2023). Generative pretrained autoregressive transformer graph neural network applied to the analysis and discovery of novel proteins. *J. Appl. Phys.* **134**. DOI:10.1063/5.0157367 DOI:10.1063/5.0157367
165. Olsen T.H., Yesiltas B., Marin F.I., et al. (2020). Anoxpepred: Using deep learning for the prediction of antioxidative properties of peptides. *Sci. Rep.* **10**:21471. DOI:10.1038/s41598-020-78319-w
166. Das P., Sercu T., Wadhawan K., et al. (2021). Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nat. Biomed. Eng.* **5**:613–623. DOI:10.1038/s41551-021-00689-x
167. Hoffman S.C., Chenthamarakshan V., Wadhawan K., et al. (2022). Optimizing molecules using efficient queries from property evaluations. *Nat. Mach. Intell.* **4**:21–31. DOI:10.1038/s42256-021-00422-y
168. Qin D., Jiao L., Wang R., et al. (2023). Prediction of antioxidant peptides using a



- quantitative structure- activity relationship predictor (AnOxPP) based on bidirectional long short-term memory neural network and interpretable amino acid descriptors. *Computers in Biol. and Med.* **154**:106591. DOI:10.1016/j.combiomed.2023.106591
169. Szymczak P., Możejko M., Grzegorzek T., et al.(2023). Discovering highly potent antimicrobial peptides with deep generative model HydrAMP. *Nat. Commun.* **14**:1453. DOI:10.1038/s41467-023-36994-z
  170. Murakami Y., Ishida S., Demizu Y., et al. (2023). Design of antimicrobial peptides containing non-proteinogenic amino acids using multi-objective Bayesian optimisation. *Digital Discov.* **2**:1347–1353. DOI:10.26434/chemrxiv-2023-cbrjc
  171. Liu Y., Zhang X., Liu Y., et al. (2023). Evolutionary multi-objective optimization in searching for various antimicrobial peptides[feature]. *IEEE Comput. Intell. Mag.* **18**:31–45. DOI:10.1109/mci.2023.3245731
  172. Tucs A., Berenger F., Yumoto A., et al. (2023). Quantum annealing designs nonhemolytic antimicrobial peptides in a discrete latent space. *ACS Med. Chem. Lett.* **14**:577–582. DOI:10.1021/acsmmedchemlett.2c00487
  173. Jain M., Bengio E., Hernandez-Garcia A., et al. (2022). Biological sequence design with GFlowNets. In *Proceedings of the 39th International Conference on Machine Learning* **162**: 9786–9801. <https://proceedings.mlr.press/v162/jain22a.html>
  174. Hesamzadeh P., Seif A., Mahmoudzadeh K., et al. (2024). *De novo* antioxidant peptide design via machine learning and DFT studies. *Sci. Rep.* **14**:6473. DOI:10.1038/s41598-024-57247-z
  175. Tyagi A., Kapoor P., Kumar R., et al. (2013). *In silico* models for designing and discovering novel anticancer peptides. *Sci. Rep.* **3**:2984. DOI:10.1038/srep02984
  176. Yi H.C., You Z.H., Zhou X., et al. (2019). ACP-DL: A deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation. *Mol. Ther. Nucleic Acids* **17**:1–9. DOI:10.1016/j.omtn.2019.04.025
  177. He W., Wang Y., Cui L., et al. (2021). Learning embedding features based on multisense-scaled attention architecture to improve the predictive performance of anticancer peptides. *Bioinformatics* **37**:4684–4693. DOI:10.1093/bioinformatics/btab560
  178. Chen J., Cheong H.H. and Siu S.W. (2021). xDeep-AcPEP: Deep learning method for anticancer peptide activity prediction based on convolutional neural network and multitask learning. *J. of Chem. Inf. and Model.* **61**:3789–3803. DOI:10.1021/acs.jcim.1c00181
  179. Deng H., Ding M., Wang Y., et al. (2023). ACP-MLC: A two-level prediction engine for identification of anticancer peptides and multi-label classification of their functional types. *Computers in Biol. and Med.* **158**:106844. DOI:10.1016/j.combiomed.2023.106844
  180. Zhong G. and Deng L. (2024). ACPScanner: Prediction of anticancer peptides by integrated machine learning methodologies. *J. of Chem. Inf. and Model.* **64**:1092–1104. DOI:10.1021/acs.jcim.3c01860
  181. Wang S. and Ma B. (2025). Anti-cancer peptides identification and activity type classification with protein sequence pre-training. *IEEE J. of Biomed. and Health Inform.* **29**:1692–1701. DOI:10.1109/JBHI.2024.3358632
  182. Grisoni F., Neuhaus C.S., Gabernet G., et al. (2018). Designing anticancer peptides by constructive machine learning. *ChemMedChem* **13**:1300–1302. DOI:10.1002/cmdc.201800204
  183. Rossetto A. and Zhou W. (2020). GANDALF: Peptide generation for drug design using sequential and structural generative adversarial networks. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* 1–10. DOI:10.1145/3388440.3412487
  184. Dou J., Vorobieva A.A., Sheffler W., et al. (2018). *De novo* design of a fluorescence-activating  $\beta$ -barrel. *Nature* **561**:485–491. DOI:10.2210/pdb6d0t/pdb
  185. Russ W.P., Figliuzzi M., Stocker C., et al. (2020). An evolution-based model for designing chorisimate mutase enzymes. *Science* **369**:440–445. DOI:10.1126/Science.aba3304
  186. Yeh A.H.W., Norn C., Kipnis Y., et al. (2023). *De novo* design of luciferases using deep learning. *Nature* **614**:774–780. DOI:10.1038/s41586-023-05696-3
  187. Zhou B., Zheng L., Wu B., et al. (2024). A conditional protein diffusion model generates artificial programmable endonuclease sequences with enhanced activity. *Cell Discov.* **10**:95. DOI:10.1101/2023.08.10.552783
  188. Cui Y., Chen Y., Sun J., et al. (2024). Computational redesign of a hydrolase for nearly complete PET depolymerization at industrially relevant high-solids loading. *Nat. Commun.* **15**:1417. DOI:10.1038/s41467-024-45662-9
  189. Crawshaw R., Crossley A.E., Johannissen L., et al. (2022). Engineering an efficient and enantioselective enzyme for the Morita–Baylis–Hillman reaction. *Nat. Chem.* **14**:313–320. DOI:10.1038/s41557-021-00833-9
  190. Lu H., Diaz D.J., Czarnicki N.J., et al. (2022). Machine learning-aided engineering of hydrolases for PET depolymerization. *Nature* **604**:662–667. DOI:10.1038/s41586-022-04599-z
  191. Lauko A., Pellock S.J., Sumida K.H., et al. (2025). Computational design of serine hydrolases. *Science* **388**:eadu2454. DOI:10.1101/2024.08.29.610411
  192. Grechishnikova D. (2021). Transformer neural network for protein-specific *de novo* drug generation as a machine translation problem. *Sci. Rep.* **11**:321. DOI:10.1101/863415
  193. Qian H., Lin C., Zhao D., et al. (2022). AlphaDrug: Protein target specific *de novo* molecular generation. *PNAS Nexus* **1**:pgac227. DOI:10.1093/pnasnexus/pgac227
  194. Chen Y., Wang Z., Wang L., et al. (2023). Deep generative model for drug design from protein target sequence. *J. of Cheminform.* **15**:38. DOI:10.1186/s13321-023-00702-2
  195. Mallick S. and Bhadra S. (2023). CDGCN: Conditional *de novo* drug generative model using graph convolution networks. In *International Conference on Research in Computational Molecular Biology* 104–119. DOI:10.1007/978-3-031-29119-7\_7
  196. Bernatavicius A., Šicho M., Janssen A.P., et al. (2024). AlphaFold meets *de novo* drug design: Leveraging structural protein information in multitarget molecular generative models. *J. of Chem. Inf. and Model.* **64**:8113–8122. DOI:10.1021/acs.jcim.4c00309
  197. Skalic M., Sabbadin D., Sattarov B., et al. (2019). From target to drug: Generative modeling for the multimodal structure-based ligand design. *Mol. Pharm.* **16**:4282–4291. DOI:10.1021/acs.molpharmaceut.9b00634
  198. Long S., Zhou Y., Dai X., et al. (2022). Zero-shot 3D drug design by sketching and generating. In *Advances in Neural Information Processing Systems* **35**: 2 3 8 9 4 – 2 3 9 0 7 . [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/96ddb813f042e8ff891b4d6f7149bb6-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/96ddb813f042e8ff891b4d6f7149bb6-Paper-Conference.pdf)
  199. Xu M., Ran T. and Chen H.(2021). *De novo* molecule design through the molecular generative model conditioned by 3D information of protein binding sites. *J. of Chem. Inf. and Model.* **61**:3240–3254. DOI:10.26434/chemrxiv.13498332.v1
  200. Krishnan S.R., Bung N., Vangala S.R., et al. (2021). *De novo* structure-based drug design using deep learning. *J. of Chem. Inf. and Model.* **62**:5100–5109. DOI:10.1021/acs.jcim.1c01319
  201. Li J.N., Yang G., Zhao P.C., et al.. (2023). CProMG: Controllable protein-oriented molecule generation with desired binding affinity and drug-like properties. *Bioinformatics* **39**:i326–i336. DOI:10.1093/bioinformatics/btad222
  202. Zhang Z., Min Y., Zheng S., et al. (2023). Molecule generation for target protein binding with structural motifs. In *The Eleventh International Conference on Learning Representations* URL <https://openreview.net/forum?id=Rq13idF0F73>
  203. Zhang Z. and Liu Q. (2023). Learning subpocket prototypes for generalizable structure-based drug design. In *Proceedings of the 40th International Conference on Machine Learning* **202**: 41382–41398. <https://proceedings.mlr.press/v202/zhang23z.html>
  204. Guan J., Qian W.W., Peng X., et al. (2023). 3D equivariant diffusion for target-aware molecule generation and affinity prediction. *arXiv abs/2303.03543*. DOI:10.48550/arXiv.2303.03543
  205. Guan J., Zhou X., Yang Y., et al. (2024). DecomDiff: Diffusion models with decomposed priors for structure-based drug design. *arXiv abs/2403.07902*. DOI:10.48550/arXiv.2403.07902
  206. Lee J., Zhong W. and Kim W.Y. (2024). NCIDiff: Non-covalent interaction-generative diffusion model for improving reliability of 3D molecule generation inside protein pocket. *arXiv abs/2405.16861*. DOI:10.48550/arXiv.2405.16861
  207. Chhabra S., Xie J. and Frank A.T. (2020). RNAPosers: Machine learning classifiers for ribonucleic acid–ligand poses. *The J. of Phys. Chem. B* **124**:4436–4445. DOI:10.1021/acs.jpcc.0c02322
  208. Stefaniak F. and Bujnicki J.M. (2021). AnnapuRNA: A scoring function for predicting RNA-small molecule binding poses. *PLoS Comput. Biol.* **17**:e1008309. DOI:10.1371/journal.pcbi.1008309
  209. Haga C.L., Yang X.D., Gheir I.S., et al. (2023). Graph neural networks for the identification of novel inhibitors of a small RNA. *SLAS Discov.* **28**:402–409. DOI:10.1016/j.slasd.2023.10.002
  210. Szulc N.A., Mackiewicz Z., Bujnicki J.M., et al. (2023). Structural interaction fingerprints and machine learning for predicting and explaining binding of small molecule ligands to RNA. *Brief. in Bioinform.* **24**:bbad187. DOI:10.1093/bib/bbad187
  211. Krishnan S.R., Roy A. and Gromiha M.M. (2024). Reliable method for predicting the binding affinity of RNA-small molecule interactions using machine learning. *Brief. in Bioinform.* **25**:bbae002. DOI:10.1093/bib/bbae002
  212. Grimberg H., Tiwari V.S., Tam B., et al. (2022). Machine learning approaches to optimize small-molecule inhibitors for RNA targeting. *J. of Cheminform.* **14**:4. DOI:10.1186/s13321-022-00583-x
  213. Segler M.H., Kogej T., Tyrchan C., et al.(2018). Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Central Sci.* **4**:120–131. DOI:10.1021/acscentsci.7b00512
  214. Zhou P., Wang J., Li C., et al.(2025). Instruction multi-constraint molecular generation using a teacher-student large language model. *BMC Biol.* **23**:105. DOI:10.21203/rs.3.rs-3845824/v1
  215. Wang J., Luo H., Qin R., et al. (2025). 3DSMILES-GPT: 3D molecular pocket-based generation with token-only large language model. *Chem. Sci.* **16**:637–648. DOI:10.1039/d4sc06864e
  216. Liao Z., Xie L., Mamitsuka H., et al. (2023). Sc2Mol: A scaffold-based two-step molecule generator with variational autoencoder and transformer. *Bioinformatics* **39**:btac814. DOI:10.1093/bioinformatics/btac814
  217. Gong H., Liu Q., Wu S., et al. (2024). Text-guided molecule generation with diffusion language model. *Proc. AAAI Conf. Artif. Intell.* **38**: 109–117. DOI:10.1609/aaai.v38i1.27761
  218. Jin W., Barzilay R. and Jaakkola T. (2018). Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning* **80**: 2323–2332. <https://proceedings.mlr.press/v80/jin18a.html>
  219. Shi C., Xu M., Zhu Z., et al. (2020). GraphAF: A flow-based autoregressive model for

- molecular graph generation. *arXiv abs/2001.09382*. DOI:10.48550/arXiv.2001.09382
220. Pirnay J., Rittig J.G., Wolf A.B., et al. (2025). GraphXForm: Graph transformer for computer-aided molecular design. *Digital Discov.* **4**:1052–1065. DOI:10.1039/d4dd00339j
221. Vignac C., Krawczuk I., Siraudin A., et al. (2022). Digress: Discrete denoising diffusion for graph generation. *arXiv abs/2209.14734*. DOI:10.48550/arXiv.2209.14734
222. Bian T., Niu Y., Chang H., et al. (2024). Hierarchical graph latent diffusion model for conditional molecule generation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* 130–140. DOI:10.1145/3627673.3679547
223. Ketata M.A., Gao N., Sommer J., et al. (2024). Lift your molecules: Molecular graph generation in latent euclidean space. *arXiv abs/2406.10513*. DOI:10.48550/arXiv.2406.10513
224. Xu M., Powers A.S., Dror R.O., et al. (2023). Geometric latent diffusion models for 3D molecule generation. In *Proceedings of the 40th International Conference on Machine Learning* **202**: 38592–38610. <https://proceedings.mlr.press/v202/xu23n.html>
225. Peng X., Guan J., Liu Q., et al. (2023). MolDiff: Addressing the atom-bond inconsistency problem in 3D molecule diffusion generation. *arXiv abs/2305.07508*. DOI:10.48550/arXiv.2305.07508
226. Chen L., Li Y., Ma Y., et al. (2025). Multiscale graph equivariant diffusion model for 3D molecule design. *Sci. Adv.* **11**:eadv0778. DOI:10.1126/sciadv.adv0778
227. Liu Y., Yang J., Ren X., et al. (2025). Multi-objective molecular design through learning latent Pareto set. In *Proceedings of the AAAI Conference on Artificial Intelligence* **39**: 19006–19014. DOI:10.1609/aaai.v39i18.34092
228. Wang J., Hsieh C.Y., Wang M., et al. (2021). Multi-constraint molecular generation based on conditional transformer, knowledge distillation and reinforcement learning. *Nat. Mach. Intell.* **3**:914–922. DOI:10.1038/s42256-021-00403-1
229. Guo J. and Schwaller P. (2024). It takes two to tango: Directly optimizing for constrained synthesizability in generative molecular design. *arXiv abs/2410.11527*. DOI:10.48550/arXiv.2410.11527
230. Yang Y., Wu Z., Yao X., et al. (2022). Exploring low-toxicity chemical space with deep learning for molecular generation. *J. of Chem. Inf. and Model.* **62**:3191–3199. DOI:10.1021/acs.jcim.2c00671
231. Imrie F., Bradley A.R., Schaar M., et al. (2020). Deep generative models for 3D linker design. *J. of Chem. Inf. and Model.* **60**:1983–1995. DOI:10.1021/acs.jcim.9b01120
232. Yang Y., Zheng S., Su S., et al. (2020). SyntaLinker: Automatic fragment linking with deep conditional transformer neural networks. *Chem. Sci.* **11**:8312–8322. DOI:10.26434/chemrxiv.12271508.v5
233. Feng Y., Yang Y., Deng W., et al. (2022). SyntaLinker-Hybrid: A deep learning approach for target specific drug design. *Artif. Intell. in the Life Sci.* **2**:100035. DOI:10.1016/j.ailsci.2022.100035
234. Huang Y., Peng X., Ma J., et al. (2022). 3DLinker: An E(3) equivariant variational autoencoder for molecular linker design. *arXiv abs/2205.07309*. DOI:10.48550/arXiv.2205.07309
235. Yu Y., Xu T., Li J., et al. (2021). A novel scalarized scaffold hopping algorithm with graph-based variational autoencoder for discovery of JAK1 inhibitors. *ACS Omega* **6**:22945–22954. DOI:10.1021/acsomega.1c03613
236. Zheng S., Lei Z., Ai H., et al. (2021). Deep scaffold hopping with multimodal transformer neural networks. *J. of Cheminform.* **13**:1–15. DOI:10.1186/s13321-021-00565-5
237. Torge J., Harris C., Mathis S.V., et al. (2023). DiffHopp: A graph diffusion model for novel drug design via scaffold hopping. *arXiv abs/2308.07416*. DOI:10.48550/arXiv.2308.07416
238. Green H., Koes D.R. and Durrant J.D. (2021). DeepFrag: A deep convolutional neural network for fragment-based lead optimization. *Chem. Sci.* **12**:8036–8047. DOI:10.1101/2021.01.07.425790
239. Imrie F., Hadfield T.E., Bradley A.R., et al. (2021). Deep generative design with 3D pharmacophoric constraints. *Chem. Sci.* **12**:14577–14589. DOI:10.1039/d1sc02436a
240. Hadfield T.E., Imrie F., Merritt A., et al. (2022). Incorporating target-specific pharmacophoric information into deep generative models for fragment elaboration. *J. of Chem. Inf. and Model.* **62**:2280–2292. DOI:10.1101/2021.10.21.465268
241. Lin H., Huang Y., Zhang O., et al. (2023). Functional-group-based diffusion for pocket-specific molecule generation and elaboration. *Adv. Neural Inf. Process. Syst.* **36**: 3 4 6 0 3 – 3 4 6 2 6 . [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/6cdd4ce9330025967dd1ed0bed3010f5-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/6cdd4ce9330025967dd1ed0bed3010f5-Paper-Conference.pdf)
242. Zang C. and Tang F. (2020). MoFlow: An invertible flow model for generating molecular graphs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 617–626. DOI:10.1145/3394486.3403104
243. You J., Liu B., Ying Z., et al. (2018). Graph convolutional policy network for goal-directed molecular graph generation. *Adv. Neural Inf. Process. Syst.* **31**. [https://proceedings.neurips.cc/paper\\_files/paper/2018/file/d60678e8f28ba9c540798ebbd31177e8-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2018/file/d60678e8f28ba9c540798ebbd31177e8-Paper.pdf)
244. Nigam A., Friederich P., Krenn M., et al. (2019). Augmenting genetic algorithms with deep neural networks for exploring the chemical space. *arXiv abs/1909.11655*. DOI:10.48550/arXiv.1909.11655
245. Carpenter E.J., Seth S., Yue N., et al. (2022). GlyNet: A multi-task neural network for predicting protein–glycan interactions. *Chem. Sci.* **13**:6669–6686. DOI:10.1101/2021.05.28.446094
246. Bojar D., Powers R.K., Camacho D.M., et al. (2021). Deep-learning resources for studying glycan-mediated host-microbe interactions. *Cell Host & Microbe* **29**:132–144. DOI:10.1016/j.chom.2020.10.004
247. Dai B., Mattox D.E. and Bailey-Kellogg C. (2021). Attention please: Modeling global and local context in glycan structure-function relationships. *bioRxiv* DOI:10.1101/2021.10.15.464532
248. Burkholz R., Quackenbush J. and Bojar D. (2021). Using graph convolutional neural networks to learn a representation for glycans. *Cell Rep.* **35**:109251. DOI:10.1101/2021.03.01.433491
249. Yang Y. and Fang Q.(2024). Prediction of glycopeptide fragment mass spectra by deep learning. *Nat. Commun.* **15**:2448. DOI:10.1038/s41467-024-46771-1
250. Urban J., Jin C., Thomsson K.A., et al. (2024). Predicting glycan structure from tandem mass spectrometry via deep learning. *Nat. Methods* **21**:1206–1215. DOI:10.1038/s41592-024-02314-6
251. Han R., Ren Z., Wang Q., et al. (2024). Synthetic biomimetic liposomes harness efferocytosis machinery for highly efficient macrophages-targeted drug delivery to alleviate inflammation. *Adv. Sci.* **11**:2308325. DOI:10.1002/adv.202308325
252. Mak H.Y., Ouyang Q., Tumanov S., et al.(2021). AGPAT2 interaction with CDP-diacylglycerol synthases promotes the flux of fatty acids through the CDP-diacylglycerol pathway. *Nat. Commun.* **12**:6877. DOI:10.1038/s41467-021-27279-4
253. Zhou Z., Tu J., Xiong X., et al. (2017). LipidCCS: Prediction of collision cross-section values for lipids with high precision to support ion mobility–mass spectrometry-based lipidomics. *Anal. Chem.* **89**:9559–9566. DOI:10.1021/acs.analchem.7b02625
254. Ou Y., Zhao J., Tripp A., et al. (2024). A deep generative model for the design of synthesizable ionizable lipids. *arXiv abs/2412.00928*. DOI:10.48550/arXiv.2412.00928
255. Soleimani S., Heydari A. and Fattahi M. (2023). Swelling prediction of calcium alginate/cellulose nanocrystal hydrogels using response surface methodology and artificial neural network. *Ind. Crops and Products* **192**:116094. DOI:10.1016/j.indcrop.2022.116094
256. Karaoglu I.C., Kebabci A.O. and Kizilel S. (2023). Optimization of gelatin methacryloyl hydrogel properties through an artificial neural network model. *ACS Appl. Mater. & Interfaces* **15**:44796–44808. DOI:10.1021/acsami.3c12207
257. Wang Y., Wallmersperger T. and Ehrenhofer A.(2023). Application of back propagation neural networks and random forest algorithms in material research of hydrogels. *PAMM* **23**:e202200278. DOI:10.1002/pamm.202200278
258. Wang Y., Wallmersperger T. and Ehrenhofer A.(2024). Prediction of hydrogel swelling states using machine learning methods. *Eng. Rep.* **6**:e12893. DOI:10.1002/eng.2.12893
259. Khavandi A., Tayebi L., Kamarian S., et al. (2023). Data-driven supervised machine learning to predict the compressive response of porous PVA/gelatin hydrogels and *in-vitro* assessments: Employing design of experiments. *Int. J. of Biol. Macromol.* **253**:126906. DOI:10.1016/j.ijbiomac.2023.126906
260. Zheng S., You H., Lam K., et al. (2024). Fracture prediction of hydrogel using machine learning and inhomogeneous multiscale network. *Adv. Theory and Simul.* **7**:2300776. DOI:10.1002/adts.202300776
261. Islamkulov M., Karakus S. and Özeroğlu C. (2023). Design artificial intelligence-based optimization and swelling behavior of novel crosslinked polymeric network hydrogels based on acrylamide-2-hydroxyethyl methacrylate and acrylamide-N-isopropylacrylamide. *Colloid and Polym. Sci.* **301**:259–272. DOI:10.1007/s00396-023-05064-7
262. Boztepe C., Künkül A. and Yüceer M. (2020). Application of artificial intelligence in modeling of the doxorubicin release behavior of pH and temperature responsive poly (NIPAAm-co-AAc)-PEG IPN hydrogel. *J. of Drug Deliv. Sci. and Technol.* **57**:101603. DOI:10.1016/j.jddst.2020.101603
263. Chen B., Dong J., Ruelas M., et al. (2022). Artificial intelligence-assisted high-throughput screening of printing conditions of hydrogel architectures for accelerated diabetic wound healing. *Adv. Funct. Mater.* **32**:2201843. DOI:10.1002/adfm.202201843
264. Liu C., Zhang X., Liu X., et al. (2024). Mechanical field guiding structure design strategy for meta-fiber reinforced hydrogel composites by deep learning. *Adv. Sci.* **11**:2310141. DOI:10.1002/adv.202310141
265. Umar M., Khan H., Hussain S., et al. (2024). Integrating DFT and machine learning for the design and optimization of sodium alginate-based hydrogel adsorbents: Efficient removal of pollutants from wastewater. *Environ. Res.* **247**:118219. DOI:10.1016/j.envres.2024.118219
266. Qiu Y., Ye H., Zhang H., et al. (2024). Machine learning-driven optimization design of hydrogel-based negative hydration expansion metamaterials. *Comput.-Aided Des.* **166**:103631. DOI:10.1016/j.cad.2023.103631
267. Cadamuro F., Piazzoni M., Gamba E., et al. (2025). Artificial intelligence tool for prediction of ECM mimics hydrogel formulations via click chemistry. *Biomater. Adv.* **175**:214323. DOI:10.1016/j.bioadv.2025.214323
268. Epa V.C., Yang J., Mei Y., et al. (2012). Modelling human embryoid body cell adhesion to a combinatorial library of polymer surfaces. *J. of Mater. Chem.*

- 22:20902–20906. DOI:10.1039/c2jm34782b
269. Kwaria R.J., Mondarte E.A.Q., Tahara H., et al. (2020). Data-driven prediction of protein adsorption on self-assembled monolayers toward material screening and design. *ACS Biomater. Sci. & Eng.* **6**:4949–4956. DOI:10.1021/acsbomaterials.0c01008
270. Rostam H.M., Fisher L.E., Hook A.L., et al. (2020). Immune-instructive polymers control macrophage phenotype and modulate the foreign body response *in vivo*. *Matter* **2**:1564–1581. DOI:10.1016/j.matt.2020.03.018
271. Wu C.T., Chang H.T., Wu C.Y., et al. (2020). Machine learning recommends affordable new Ti alloy with bone-like modulus. *Mater. Today* **34**:41–50. DOI:10.1016/j.mattod.2019.08.008
272. Chang Y.J., Jui C.Y., Lee W.J., et al. (2019). Prediction of the composition and hardness of high-entropy alloys by machine learning. *JOM* **71**:3433–3442. DOI:10.1007/s11837-019-03704-4
273. Spurrell C.H., Barozzi I., Kosicki M., et al. (2022). Genome-wide fetalization of enhancer architecture in heart disease. *Cell Rep.* **40**. DOI:10.1016/j.celrep.2022.111400
274. Taskiran I.I., Spanier K.I., Dickmanken H., et al. (2024). Cell-type-directed design of synthetic enhancers. *Nature* **626**:212–220. DOI:10.1101/2022.07.26.501466
275. Stark H., Jing B., Wang C., et al. (2024). Dirichlet flow matching with applications to DNA sequence design. *arXiv abs/2402.05841*. DOI:10.48550/arXiv.2402.05841
276. Huang D., Petrykowska H.M., Miller B.F., et al. (2019). Identification of human silencers by correlating cross-tissue epigenetic profiles and gene expression. *Genome Res.* **29**:657–667. DOI:10.1101/gr.247007.118
277. Doni Jayavelu N., Jajodia A., Mishra A., et al. (2020). Candidate silencer elements for the human and mouse genomes. *Nat. Commun.* **11**:1061. DOI:10.1038/s41467-020-14853-5
278. Santos A.M., Hogan M.S., Luther R.D., et al. (2022). Genomic context sensitivity of insulator function. *Genome Res.* **32**:425–436. DOI:10.1101/2021.05.16.444030
279. De Goede O., Ferraro N., Nachun D., et al. (2019). Long non-coding RNA gene regulation and trait associations across human tissues. *bioRxiv* DOI:10.1101/793091
280. Gracia B. and Russell R. (2014). RNA catalytic activity as a probe of chaperone-mediated RNA folding. *Waldsich C. (ed.). RNA Folding: Methods and Protocols* (Humana Press), pp: 225–237. DOI:10.1007/978-1-62703-667-2\_13
281. Maas S. and Rich A. (2000). Changing genetic information through RNA editing. *Bioessays* **22**:790–802. DOI:10.1002/1521-1878(200009)22:9<790::aid-bies4>3.3.co;2-s
282. Ching T., Himmelstein D.S., Beaulieu-Jones B.K., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *J. of the R. Soc. Interface* **15**:20170387. DOI:10.1098/rsif.2017.0387
283. Eraslan G., Avsec Z., Gagneur J., et al. (2019). Deep learning: New computational modelling techniques for genomics. *Nat. Rev. Genet.* **20**:389–403. DOI:10.1038/s41576-019-0122-6
284. Alipanahi B., Delong A., Weirauch M.T., et al. (2015). Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**:831–838. DOI:10.1038/nbt.3300
285. Senior A.W., Evans R., Jumper J., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature* **577**:706–710. DOI:10.1038/s41586-019-1923-7
286. Calonaci N., Jones A., Cutarello F., et al. (2020). Machine learning a model for RNA structure prediction. *NAR Genomics and Bioinform.* **2**:lqaa090. DOI:10.1093/nargab/lqaa090
287. Chuai G., Ma H., Yan J., et al. (2018). DeepCRISPR: Optimized CRISPR guide RNA design by deep learning. *Genome Biol.* **19**:1–18. DOI:10.1186/s13059-018-1459-4
288. Zhou T., Dai N., Li S., et al. (2023). RNA design via structure-aware multifrontier ensemble optimization. *Bioinformatics* **39**:i563–i571. DOI:10.1093/bioinformatics/btad252
289. Costa F., Kohvaei P. and Kleinkauf R. (2016). RNAsynth: Constraints learning for RNA inverse folding. In *European Symposium on Artificial Neural Networks* <https://www.esann.org/sites/default/files/proceedings/legacy/es2016-137.pdf>
290. Zhang H., Zhang L., Lin A., et al. (2023). Algorithm for optimized mRNA design improves stability and immunogenicity. *Nature* **621**:396–403. DOI:10.1038/s41586-023-06127-z
291. Tang X., Huo M., Chen Y., et al. (2024). A novel deep generative model for mRNA vaccine development: Designing 5' UTRs with m1-methyl-pseudouridine modification. *Acta Pharm. Sin. B* **14**:1814–1826. DOI:10.1016/j.apsb.2023.11.003
292. Hofacker I.L., Fontana W., Stadler P.F., et al. (1994). Fast folding and comparison of RNA secondary structures. *Monatsh. für Chem.* **125**:167–167. DOI:10.1007/bf00818163
293. Busch A. and Backofen R. (2006). INFO-RNA—a fast approach to inverse RNA folding. *Bioinformatics* **22**:1823–1831. DOI:10.1093/bioinformatics/btl194
294. Nomura Y. and Yokobayashi Y. (2019). Systematic minimization of RNA ligase ribozyme through large-scale design-synthesis-sequence cycles. *Nucleic Acids Res.* **47**:8950–8960. DOI:10.1093/nar/gkz729
295. Ning H., Liu G., Li L., et al. (2023). Rational design of microRNA-responsive switch for programmable translational control in mammalian cells. *Nat. Commun.* **14**:7193. DOI:10.1038/s41467-023-43065-w
296. Nieuwkoop T., Terlouw B.R., Stevens K.G., et al. (2023). Revealing determinants of translation efficiency via whole-gene codon randomization and machine learning. *Nucleic Acids Res.* **51**:2363–2376. DOI:10.1093/nar/gkad035
297. Castillo-Hair S., Fedak S., Wang B., et al. (2024). Optimizing 5'UTRs for mRNA-delivered gene editing using deep learning. *Nat. Commun.* **15**:5284. DOI:10.1038/s41467-024-49508-2
298. Ren Z., Jiang L., Di Y., et al. (2024). CodonBERT: A BERT-based architecture tailored for codon optimization using the cross-attention mechanism. *Bioinformatics* **40**:btac330. DOI:10.1093/bioinformatics/btae330
299. Constant D.A., Gutierrez J.M., Sastry A.V., et al. (2023). Deep learning-based codon optimization with large-scale synonymous variant datasets enables generalized tunable protein expression. *bioRxiv* DOI:10.1101/2023.02.11.528149
300. Gong H., Wen J., Luo R., et al. (2023). Integrated mRNA sequence optimization using deep learning. *Brief. in Bioinform.* **24**:bbad001. DOI:10.1093/bib/bbad001
301. Sumi S., Hamada M. and Saito H. (2024). Deep generative design of RNA family sequences. *Nat. Methods* **21**:435–443. DOI:10.1038/s41592-023-02148-8
302. Zuker M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**:3406–3415. DOI:10.1093/nar/gkg595
303. Markham N.R. and Zuker M. (2008). UNAFold: Software for nucleic acid folding and hybridization. *Keith J.M. (ed.). Bioinformatics: Structure, function and applications* (Humana Press), pp: 3–31. DOI:10.1007/978-1-60327-429-6\_1
304. Sato K., Akiyama M. and Sakakibara Y. (2021). RNA secondary structure prediction using deep learning with thermodynamic integration. *Nat. Commun.* **12**:941. DOI:10.1101/2020.08.10.244442
305. Fu L., Cao Y., Wu J., et al. (2022). Ufold: Fast and accurate RNA secondary structure prediction with deep learning. *Nucleic Acids Res.* **50**:e14–e14. DOI:10.1101/2020.08.17.254896
306. Shen T., Hu Z., Peng Z., et al. (2022). E2efold-3D: End-to-end deep learning method for accurate *de novo* RNA 3D structure prediction. *arXiv abs/2207.01586*. DOI:10.48550/arXiv.2207.01586
307. Justyna M., Zirlab C., Antozak M., et al. (2025). Graph neural network and diffusion model for modeling rna interatomic interactions. *Bioinformatics* **41**:btaf515. DOI:10.1093/bioinformatics/btaf515
308. Nori D. and Jin W. (2024). RNAFlow: RNA structure & sequence design via inverse folding-based flow matching. *arXiv abs/2405.18768*. DOI:10.48550/arXiv.2405.18768
309. Xia Y., Liang Z., Du X., et al. (2025). Design of function-regulating RNA via deep learning and AlphaFold 3. *Brief. in Bioinform.* **26**:bbaf419. DOI:10.1093/bib/bbaf419
310. Ghosh A., Larrondo-Petrie M.M. and Pavlovic M. (2023). Revolutionizing vaccine development for COVID-19: A review of AI-based approaches. *Information* **14**:665. DOI:10.20944/preprints202311.0656.v1
311. Imani S., Li X., Chen K., et al. (2025). Computational biology and artificial intelligence in mRNA vaccine design for cancer immunotherapy. *Front. in Cell. and Infect. Microbiol.* **14**:1501010. DOI:10.3389/fcimb.2024.1501010
312. Zhu Y., Zhu L., Wang X., et al. (2022). RNA-based therapeutics: An overview and prospectus. *Cell Death & Disease* **13**:644. DOI:10.1038/s41419-022-05075-2
313. Paunovska K., Loughrey D. and Dahlman J.E. (2022). Drug delivery systems for RNA therapeutics. *Nat. Rev. Genet.* **23**:265–280. DOI:10.1038/s41576-021-00439-4
314. Murdoch B. (2021). Privacy and artificial intelligence: Challenges for protecting health information in a new era. *BMC Med. Ethics* **22**:122. DOI:10.1186/s12910-021-00687-3
315. Norori N., Hu Q., Aellen F.M., et al. (2021). Addressing bias in big data and AI for health care: A call for open science. *Patterns* **2**:100347. DOI:10.1016/j.patter.2021.100347 DOI:10.1016/j.patter.2021.100347
316. Pannu J., Bloomfield D., MacKnight R., et al. (2025). Dual-use capabilities of concern of biological AI models. *PLoS Comput. Biol.* **21**:e1012975. DOI:10.1371/journal.pcbi.1012975
317. Ferruz N. and Höcker B. (2022). Controllable protein design with language models. *Nat. Mach. Intell.* **4**:521–532. DOI:10.1038/s42256-022-00499-z
318. Yang C., Sesterhenn F., Bonet J., et al. (2021). Bottom-up *de novo* design of functional proteins with complex structural features. *Nat. Chem. Biol.* **17**:492–500. DOI:10.1038/s41589-020-00699-x
319. Sesterhenn F., Yang C., Bonet J., et al. (2020). *De novo* protein design enables the precise induction of RSV-neutralizing antibodies. *Science* **368**:eaay5051. DOI:10.1126/Science.aay5051
320. Zhu J., Avakyan N., Kakkis A., et al. (2021). Protein assembly by design. *Chem. Rev.* **121**:13701–13796. DOI:10.1021/acs.chemrev.1c00308
321. Miserez A., Yu J. and Mohammadi P. (2023). Protein-based biological materials: Molecular design and artificial production. *Chem. Rev.* **123**:2049–2111. DOI:10.1021/acs.chemrev.2c00621
322. Linsky T.W., Vergara R., Codina N., et al. (2020). *De novo* design of potent and resilient hACE2 decoys to neutralize SARS-CoV-2. *Science* **370**:1208–1214. DOI:10.1126/Science.abe0075
323. Anand N. and Achim T. (2022). Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *arXiv abs/2205.15019*. DOI:10.48550/arXiv.2205.15019
324. Luo S., Su Y., Peng X., et al. (2022). Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. *Adv.*



- Neural Inf. Process. Syst. **35**: 9754–9767. DOI:10.1101/2022.07.10.499510
325. Yim J., Trippe B.L., De Bortoli V., et al. (2023). Se(3) diffusion model with application to protein backbone generation. *arXiv abs/2302.02277*. DOI:10.48550/arXiv.2302.02277
326. Fu C., Yan K., Wang L., et al. (2024). A latent diffusion model for protein structure generation. In *Proceedings of the Second Learning on Graphs Conference* **231**: 29: 1–29: 17. <https://proceedings.mlr.press/v231/fu24a.html>
327. Wu L., Trippe B., Naesseth C., et al. (2023). Practical and asymptotically exact conditional sampling in diffusion models. *Adv. Neural Inf. Process. Syst.* **36**: 3 1 3 7 2 – 3 1 4 0 3 . [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/63e8bc7bbf1cfea36d1d1b6538aeece5-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/63e8bc7bbf1cfea36d1d1b6538aeece5-Paper-Conference.pdf)
328. Alamdari S., Thakkar N., Berg R., et al. (2023). Protein generation with evolutionary diffusion: Sequence is all you need. *bioRxiv* DOI:10.1101/2023.09.11.556673
329. Janson G., Valdes-Garcia G., Heo L., et al. (2023). Direct generation of protein conformational ensembles via machine learning. *Nat. Commun.* **14**:774. DOI:10.1038/s41467-023-36443-x
330. Lee J.S., Kim J. and Kim P.M. (2023). Score-based generative modeling for *de novo* protein design. *Nat. Comput. Sci.* **3**:382–392. DOI:10.1038/s43588-023-00440-3
331. Trippe B.L., Yim J., Tischler D., et al. (2022). Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. *arXiv abs/2206.04119*. DOI:10.48550/arXiv.2206.04119
332. Wu K.E., Yang K.K., Berg R., et al. (2024). Protein structure generation via folding diffusion. *Nat. Commun.* **15**:1059. DOI:10.1038/s41467-024-45051-2
333. Lin Y. and AlQuraishi M. (2023). Generating novel, designable, and diverse protein structures by equivalently diffusing oriented residue clouds. *arXiv abs/2301.12485*. DOI:10.48550/arXiv.2301.12485
334. Frentzas S., Ahern E.S., Weickhardt A.J., et al. (2023). A phase 1/2 study of AU-007, a monoclonal antibody (mAb) that binds to IL-2 and inhibits CD25 binding, in patients with advanced solid tumors: Interim results. *J. of Clin. Oncol.* **41**:e14507–e14507. DOI:10.1200/JCO.2023.41.16\_suppl.e14507
335. Rasco D.W., McKean M., Haydon A.M., et al. (2024). Updated results of a phase 1/2 study of AU-007, a monoclonal antibody (mAb) that binds to IL-2 and inhibits CD25 binding, in patients with advanced solid tumors. *J. of Clin. Oncol.* **42**:2527–2527. DOI:10.1200/JCO.2024.42.16\_suppl.2527
336. Fjell C.D., Hiss J.A., Hancock R.E., et al. (2012). Designing antimicrobial peptides: Form follows function. *Nat. Rev. Drug Discov.* **11**:37–51. DOI:10.1038/nrd3591
337. Yan J., Cai J., Zhang B., et al. (2022). Recent progress in the discovery and design of antimicrobial peptides using traditional machine learning and deep learning. *Antibiotics* **11**:1451. DOI:10.3390/antibiotics11101451
338. Silva O.N., Torres M.D., Cao J., et al. (2020). Repurposing a peptide toxin from wasp venom into antileishmaniasis with dual antimicrobial and immunomodulatory properties. *Proc. Natl. Acad. Sci.* **117**:26936–26945. DOI:10.1073/pnas.2012379117
339. Johnson S.R., Fu X., Vikander S., et al. (2025). Computational scoring and experimental evaluation of enzymes generated by neural networks. *Nat. Biotechnol.* **43**:396–405. DOI:10.1101/2023.03.04.531015
340. Munsamy G., Illanes-Vicioso R., Funcillo S., et al. (2024). Conditional language models enable the efficient design of proficient enzymes. *bioRxiv* DOI:10.1101/2024.05.03.592223
341. Barghout R.A., Xu Z., Betala S., et al. (2023). Advances in generative modeling methods and datasets to design novel enzymes for renewable chemicals and fuels. *Curr. Opin. in Biotechnol.* **84**:103007. DOI:10.1016/j.copbio.2023.103007
342. Zimmerman L., Alon N., Levin I., et al. (2024). Context-dependent design of induced-fit enzymes using deep learning generates well-expressed, thermally stable and active enzymes. *Proc. Natl. Acad. Sci.* **121**:e2313809121. DOI:10.1101/2023.07.27.550799
343. Wang Q., Liu X., Zhang H., et al. (2024). Cytochrome P450 enzyme design by constraining the catalytic pocket in a diffusion model. *Res.* **7**:0413. DOI:10.1101/2024.01.08.574609
344. Cao L., Coventry B., Goreshnik I., et al. (2022). Design of protein-binding proteins from the target structure alone. *Nature* **605**:551–560. DOI:10.1038/s41586-022-04654-9
345. Jansen R., Schumann G., Baadenhuijsen H., et al. (2006). Trueness verification and traceability assessment of results from commercial systems for measurement of six enzyme activities in serum: An international study in the EC4 framework of the Calibration 2000 project. *Clinica Chimica Acta* **368**:160–167. DOI:10.1016/j.cca.2005.12.033
346. Broom A., Rakotoharisoa R.V., Thompson M.C., et al. (2020). Ensemble-based enzyme design can recapitulate the effects of laboratory directed evolution in silico. *Nat. Commun.* **11**:4808. DOI:10.1038/s41467-020-18619-x
347. Wang Z., Xie D., Wu D., et al. (2025). Robust enzyme discovery and engineering with deep learning using CataPro. *Nat. Commun.* **16**:2736. DOI:10.1038/s41467-025-58038-4
348. Hunter P. (2024). Security challenges by AI-assisted protein design: The ability to design proteins in silico could pose a new threat for biosecurity and biosafety. *EMBO Rep.* **25**:2168–2171. DOI:10.1038/s44319-024-00124-7
349. Stuart L.M., Bright R.A. and Horvitz E. (2024). AI-enabled protein design: A strategic asset for global health and biosecurity. *NAM Perspect.* **2024**:10–31478. DOI:10.31478/202410d
350. Baker D. and Church G. (2024). Protein design meets biosecurity. *Science* **383**:349–349. DOI:10.1126/science.ado1671
351. Fischer T., Gazzola S. and Riedl R. (2019). Approaching target selectivity by *de novo* drug design. *Expert Opin. on Drug Discov.* **14**:791–803. DOI:10.1080/17460441.2019.1615435
352. Schneider G. and Fechner U. (2005). Computer-based *de novo* design of drug-like molecules. *Nat. Rev. Drug Discov.* **4**:649–663. DOI:10.1038/nrd1799
353. Wu K., Karapetyan E., Schloss J., et al. (2023). Advancements in small molecule drug design: A structural perspective. *Drug Discov. Today* **28**:103730. DOI:10.1016/j.drudis.2023.103730
354. Thomas M., Bender A. and Graaf C. (2023). Integrating structure-based approaches in generative molecular design. *Curr. Opin. in Struct. Biol.* **79**:102559. DOI:10.1016/j.sbi.2023.102559
355. Yu A.M., Choi Y.H. and Tu M.J. (2020). RNA drugs and RNA targets for small molecules: Principles, progress, and challenges. *Pharmacol. Rev.* **72**:862–898. DOI:10.1124/pr.120.019554
356. Santos R., Ursu O., Gaulton A., et al. (2017). A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.* **16**:19–34. DOI:10.1038/nrd.2016.230
357. Usmani S.S., Bedi G., Samuel J.S., et al. (2017). THPdb: Database of FDA-approved peptide and protein therapeutics. *PLoS One* **12**:e0181748. DOI:10.1371/journal.pone.0181748
358. Rock B.M. and Foti R.S. (2019). Pharmacokinetic and drug metabolism properties of novel therapeutic modalities. *Drug Metab. and Disposition* **47**:1097–1099. DOI:10.1124/dmd.119.088708
359. Yin W. and Rogge M. (2019). Targeting RNA: A transformative therapeutic strategy. *Clin. and Transl. Sci.* **12**:98–112. DOI:10.1111/cts.12624
360. Sheng J., Gan J. and Huang Z. (2013). Structure-based DNA-targeting strategies with small molecule ligands for drug discovery. *Med. Res. Rev.* **33**:1119–1173. DOI:10.1002/med.21278
361. Xu Z., Ren F., Wang P., et al. (2025). A generative AI-discovered TNiK inhibitor for idiopathic pulmonary fibrosis: A randomized phase 2a trial. *Nat. Med.* **31**:2602–2610. DOI:10.1038/s41591-025-03743-2
362. Warner K.D., Hajdin C.E. and Weeks K.M. (2018). Principles for targeting RNA with drug-like small molecules. *Nat. Rev. Drug Discov.* **17**:547–558. DOI:10.1038/nrd.2018.93
363. Childs-Disney J.L., Yang X., Gibaut Q.M., et al. (2022). Targeting RNA structures with small molecules. *Nat. Rev. Drug Discov.* **21**:736–762. DOI:10.1038/s41573-022-00521-4
364. Watanabe K. and Seki N. (2024). Biology and development of DNA-targeted drugs, focusing on synthetic lethality, DNA repair, and epigenetic modifications for cancer: A review. *Int. J. Mol. Sci.* **25**:752. DOI:10.3390/ijms25020752
365. Morishita E.C. and Nakamura S. (2024). Recent applications of artificial intelligence in RNA-targeted small molecule drug discovery. *Expert Opin. on Drug Discov.* **19**:415–431. DOI:10.1080/17460441.2024.2313455
366. Rizvi N.F., Santa Maria Jr J.P., Nahvi A., et al. (2020). Targeting RNA with small molecules: Identification of selective, RNA-binding small molecules occupying drug-like chemical space. *SLAS Discov.: Advancing the Sci. of Drug Discov.* **25**:384–396. DOI:10.1177/2472555219885373
367. Zhang O., Lin H., Zhang H., et al. (2024). Deep lead optimization: Leveraging generative AI for structural modification. *J. of the Am. Chem. Soc.* **146**:31357–31370. DOI:10.1021/jacs.4c11686
368. Mészáros B., Járvas G., Kun R., et al. (2020). Machine learning based analysis of human serum *n*-glycome alterations to follow up lung tumor surgery. *Cancers* **12**:3700. DOI:10.3390/cancers12123700
369. Huang C., Fang M., Feng H., et al. (2021). N-glycan fingerprint predicts alpha-fetoprotein negative hepatocellular carcinoma: A large-scale multicenter study. *Int. J. of Cancer* **149**:717–727. DOI:10.1002/ijc.33564
370. Remo E., Markus O., Giorgio B., et al. (2024). Leveraging machine learning to streamline the development of liposomal drug delivery systems. *J. of Controlled Release* **376**:1–12. DOI:10.1016/j.jconrel.2024.10.065
371. Zhu X., et al. (2025). Machine learning-based glycolipid metabolism gene signature predicts prognosis and immune landscape in oesophageal squamous cell carcinoma. *J. of Cell. and Mol. Med.* **29**:e70434. DOI:10.1111/jcmm.70434
372. Ratner B.D. (2019). Biomaterials: Been there, done that, and evolving into the future. *Annu. Rev. of Biomed. Eng.* **21**:171–191. DOI:10.1146/annurev-bioeng-062117-120940
373. Li J. and Mooney D.J. (2016). Designing hydrogels for controlled drug delivery. *Nat. Rev. Mater.* **1**:1–17. DOI:10.1038/natrevmats.2016.71
374. Suwardi A., Wang F., Xue K., et al. (2022). Machine learning-driven biomaterials evolution. *Adv. Mater.* **34**:2102703. DOI:10.1002/adma.202102703
375. Negut I. and Bitá B. (2023). Exploring the potential of artificial intelligence for hydrogel development—a short review. *Gels* **9**:845. DOI:10.3390/gels9110845
376. Xu J., Cao B., Yang H., et al. (2024). Machine learning-assisted design of low elastic modulus  $\beta$ -type medical titanium alloys and experimental validation. *Comput. Mater. Sci.* **238**:112927. DOI:10.1016/j.commatsci.2024.112902
377. Horikawa S., Suzuki K., Motojima K., et al. (2024). Material design of porous hydroxyapatite ceramics via inverse analysis of an estimation model for bone-



- forming ability based on machine learning and experimental validation of biological hard tissue responses. *Mater.* **17**:571. DOI:10.3390/ma17030571
378. Gangwal A., Ansari A., Ahmad I., et al. (2024). Current strategies to address data scarcity in artificial intelligence-based drug discovery: A comprehensive review. *Computers in Biol. and Med.* **179**:108734. DOI:10.1016/j.compbiomed.2024.108734
  379. Wu T., Zhou M., Zou J., et al. (2024). AI-guided few-shot inverse design of HDP-mimicking polymers against drug-resistant bacteria. *Nat. Commun.* **15**:6288. DOI:10.1038/s41467-024-50533-4
  380. Cai F., Hanna K., Zhu T., et al. (2024). A foundation model for chemical design and property prediction. *arXiv abs/2410.21422*. DOI:10.48550/arXiv.2410.21422
  381. Ross J., Belgodere B., Chenthamarakshan V., et al. (2022). Large-scale chemical language representations capture molecular structure and properties. *Nat. Mach. Intell.* **4**:1256–1264. DOI:10.1038/s42256-022-00580-7
  382. Kim S., Nam J., Yu S., et al. (2024). Data-efficient molecular generation with hierarchical textual inversion. *arXiv abs/2405.02845*. DOI:10.48550/arXiv.2405.02845
  383. Méndez-Lucio O., Nicolaou C.A. and Earnshaw B. (2024). MolE: A foundation model for molecular graphs using disentangled attention. *Nat. Commun.* **15**:9431. DOI:10.1038/s41467-024-53751-y

## FUNDING AND ACKNOWLEDGMENTS

This work is partially supported by the National Natural Science Foundation of China [62506351, 62372375, 92470204], National Key R&D Program of China [2023YFC2604400], Natural Science Foundation of Beijing Municipality [4254089], China Postdoctoral Science Foundation [2023M743565, GZC20232736] and Decentralised Gap funding [I23D1AG081] from A\*STAR, Singapore. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

## AUTHOR CONTRIBUTIONS

Guang Yang conceived the study, coordinated all contributors, wrote the RNA section, performed final editing and proofreading, integrated all sections, and handled manuscript submission. Meng Xiao wrote the generative AI section and created related figures. Jianing Li wrote the protein structure-based and ligand-based *de novo* design sections of small molecules, created the graphical abstract and main figures, and refined all illustrations. Qiong Wang wrote the peptide section and created figures for the protein chapter. Lianlian Wu wrote the nucleic acid-based design and molecular modification sections of small molecules. Zhengshan Chen wrote the antibody section. Peng-Cheng Zhao and Qinglong Wang wrote the glycan and lipid sections, with Qinglong Wang also creating figures. Haoyang Wang wrote the biomaterials section and created related figures. Bing Wen wrote the DNA section and created preliminary figures. Chao Song wrote the functional protein and enzyme sections. Wenxin Xu contributed to writing the glycan and lipid sections. Xiongfei He and Xiao Zhang contributed to writing the biomaterials section. All corresponding authors supervised the project and provided critical revisions. All authors contributed to the manuscript and approved the final version.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

## DATA AND CODE AVAILABILITY

Data availability is not applicable to this article as no new data were created or analyzed in this study.