Acta Pharmaceutica Sinica B 2026;16(7):3996−4023

Chinese Pharmaceutical Association
Institute of Materia Medica, Chinese Academy of Medical Sciences

Acta Pharmaceutica Sinica B

www.elsevier.com/locate/apsb
www.sciencedirect.com

REVIEW

Artificial intelligence in biologic drug discovery: A review of methodological evolution and therapeutic applications

Jianxin Tang[a,†], Daohong Gong[a,†], Honglin Li[a,*], Shiliang Li[a,b,*]

[a]*Innovation Center for AI and Drug Discovery, School of Pharmacy, East China Normal University, Shanghai 200062, China*
[b]*Department of Pain Management, HuaDong Hospital Affiliated to Fudan University, Shanghai 200040, China*

Received 30 June 2025; received in revised form 23 January 2026; accepted 27 January 2026

KEY WORDS

Artificial intelligence;
Biologic drugs;
De novo design;
Drug optimization;
Drug delivery system;
Nucleic acid therapeutics;
Machine learning;
Generative models

Abstract Biologic drugs, primarily comprising proteins and nucleic acids, have emerged as powerful therapeutic modalities; however, their discovery and optimization are often hindered by their inherent complexity. The advent of artificial intelligence (AI), particularly deep learning, is catalyzing a paradigm shift in this field, transitioning it from a process reliant on serendipity and laborious experimentation to a data-driven engineering discipline. This review systematically charts the co-evolution of AI methodologies and their transformative applications across the modern biologic drug development pipeline. We first outline AI's methodological progression, from language models deciphering biological sequence grammar to structure prediction models like AlphaFold making macromolecular folds computationally accessible, and finally to generative models enabling *de novo* molecular creation. We then explore the practical impact of these technologies in two core phases: the *de novo* design of novel biologics with bespoke functions and the subsequent multi-parameter engineering and optimization of these candidates for clinical viability. While the potential is immense, significant strategic challenges remain, including the need to build a new AI-native experimental ecosystem and bridge the profound complexity gap between molecular-level predictions and systemic *in vivo* outcomes. Overcoming these obstacles will usher in a new era of AI-driven, automated closed-loop drug discovery.

This article is part of special issue entitled Machine Learning in Drug Discovery.
*Corresponding authors.
E-mail addresses: hlli@hsc.ecnu.edu.cn (Honglin Li), slli@hsc.ecnu.edu.cn (Shiliang Li).
[†]These authors made equal contributions to this work.
Peer review under the responsibility of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences.

https://doi.org/10.1016/j.apsb.2026.01.039

1. Introduction

The therapeutic landscape has undergone a profound and significant evolution over the past three decades, marked by the ascendancy of biologic drugs[1,2]. Biologics, a class of therapeutics principally composed of large, complex molecules such as proteins and nucleic acids, have emerged from the periphery of pharmaceutical research to become a central pillar of modern medicine[3]. Unlike traditional small-molecule drugs, which typically function by inhibiting enzyme active sites or receptor pockets, macromolecules operate through highly specific and often multifaceted mechanisms[4,5]. Monoclonal antibodies, for instance, can precisely target cell surface receptors with exquisite affinity, flag cancer cells for destruction by the immune system, or neutralize circulating cytokines with unparalleled specificity[6]. Therapeutic enzymes[7,8] can replace deficient endogenous proteins in metabolic disorder[9]. At the same time, nucleic acid-based therapies, such as small interfering RNA (siRNA) and messenger RNA (mRNA), can modulate gene expression at its most fundamental level, silencing disease-causing genes or instructing cells to produce their own therapeutic proteins[10,11]. This functional versatility has unlocked treatments for a vast spectrum of complex human diseases, from oncological and autoimmune disorders to rare genetic conditions and infectious diseases, for which conventional chemical entities offered limited solutions[5].

However, the very source of their therapeutic power—their intricate, high-molecular-weight structures and complex biological functions—simultaneously constitutes their greatest developmental challenge. The design space for a macromolecule is staggeringly vast; a modest 100-residue protein possesses a theoretical sequence space of 20^{100} possibilities, a number far exceeding the number of atoms in the known universe. This combinatorial explosion renders exhaustive experimental exploration a physical impossibility. Consequently, the discovery and engineering of biologics have historically relied on a paradigm that is both laborious and fraught with uncertainty. The process often begins with either serendipitous discovery through large-scale screening of natural or synthetic libraries[12] or painstaking rational design based on limited structural information and biophysical intuition[13]. Subsequent optimization to enhance properties like affinity, stability, and solubility involves iterative, low-throughput site-directed mutagenesis—a process characterized by high costs, long timelines, and a limited ability to navigate the complex, epistatic interactions that govern a protein's fitness landscape[14,15]. This established paradigm is now being fundamentally disrupted by the confluence of two powerful currents: the deluge of biological data from high-throughput sequencing and structural biology, and the concurrent maturation of artificial intelligence (AI), especially deep learning[16–18]. AI models, particularly neural networks with deep architectures, have demonstrated an extraordinary capacity to learn complex, nonlinear patterns from high-dimensional data[19]. This makes them uniquely suited to tackling the central challenge in biologic science: deciphering the intricate mapping from a one-dimensional sequence to a three-dimensional structure and, ultimately, to a specific biological function. The integration of AI is therefore not just an incremental improvement in accelerating data analysis; it also transforms macromolecular research from an empirical endeavor to a data-driven discipline. AI is becoming a creative technology in the design process, capable of intelligently navigating the vast sequence space with the biological principles learned[20–23].

This review provides a structured overview of this transformation. We first introduce the foundational AI methodologies, from sequence and structure prediction to generative modeling (Section 3). We then systematically explore their applications in two core drug discovery phases: the *de novo* design of novel biologics (Section 4) and the subsequent multi-parameter engineering and optimization of these candidates (Section 5). Following this, we examine the critical role of AI in developing advanced drug delivery systems (Section 6). Finally, we synthesize the overarching challenges and strategic frontiers that will shape the future of this rapidly evolving field (Section 7), concluding with a perspective on the path toward a fully integrated, AI-driven discovery paradigm.

2. Brief introduction to biologic drugs and related applications

Biologic drugs have become a cornerstone of modern medicine. These powerful therapeutics are fundamentally classified based on their distinct molecular compositions and mechanisms of action. As shown in Fig. 1A, they fall into two broad categories: protein drugs and nucleic acid drugs. Protein drugs, such as antibodies, peptides, and enzymes, are assembled from 20 natural amino acids and act mainly through direct molecular interactions, including antigen binding, receptor modulation, or enzymatic catalysis. In contrast, nucleic acid drugs, including mRNA, siRNA, and ASO, are built from four nucleotide bases and function primarily at the genetic level by encoding proteins, silencing target genes, or modulating RNA processing. This inherent difference in composition and mechanism is critical for guiding drug design, therapeutic optimization, and delivery strategy development. Fig. 1B illustrates the high specificity and efficacy of biologic drugs confer significant advantages, supporting their broad application across diverse therapeutic areas, including vaccine development, gene therapy, cancer treatment, autoimmune and cardiovascular disorders, and rare diseases.

2.1. Protein drugs

Protein-based drugs represent a foundational class of biologic therapeutics, leveraging the inherent specificity and diverse biological functions of naturally occurring proteins or their engineered counterparts. Their therapeutic efficacy stems from their ability to precisely engage and modulate specific biological pathways, making them indispensable across a wide spectrum of diseases (Table 1). These agents are systematically categorized

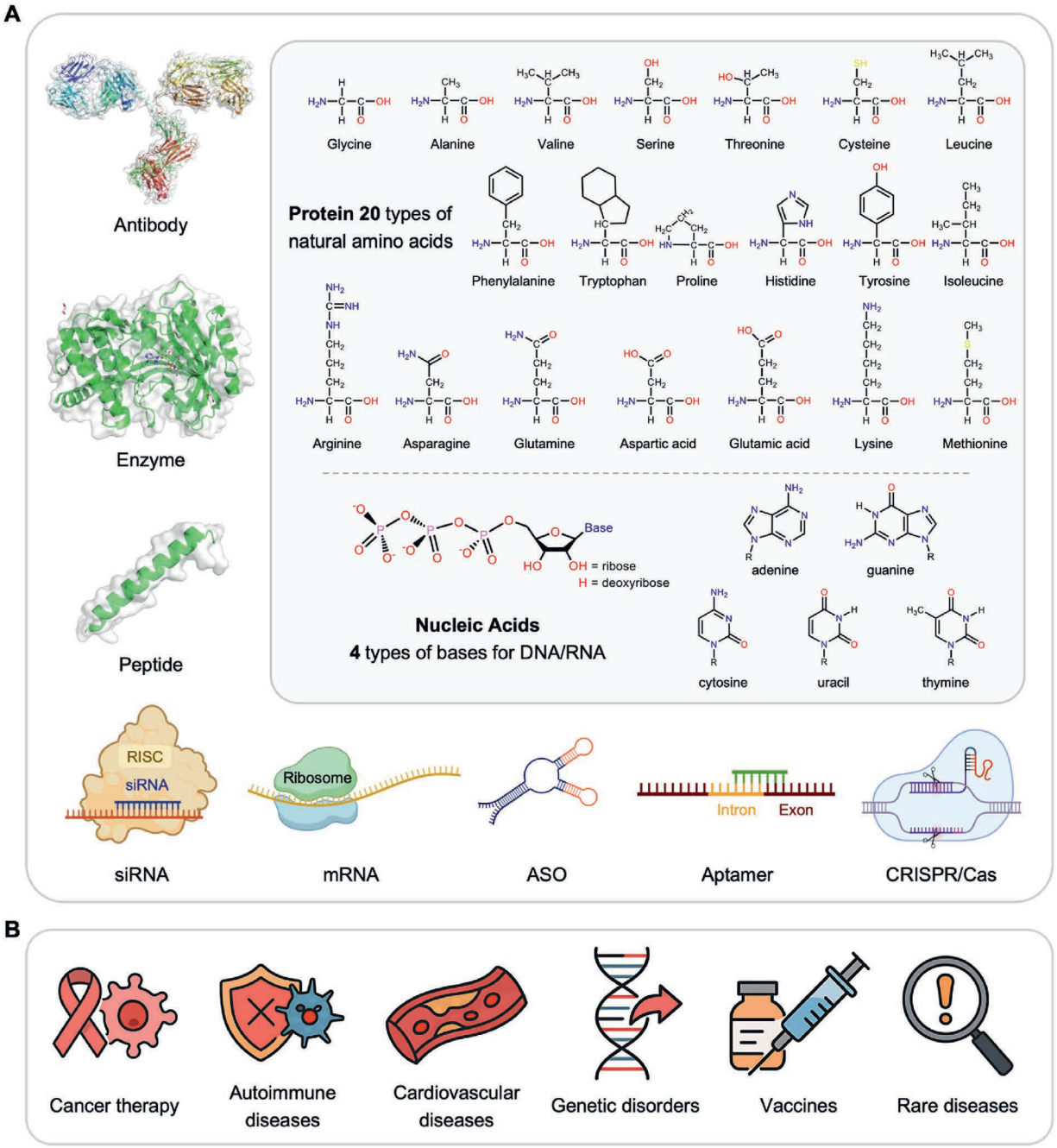

Figure 1 Compositions and therapeutic applications of biologic drugs. (A) The main classes of biologic drugs: protein drugs, consisting of 20 types of natural amino acids, and nucleic acid drugs, which are built from 4 types of bases for DNA/RNA. (B) The therapeutic applications of biologic drugs, including cancer therapy, vaccines, genetic disorders therapy, treatments for autoimmune, cardiovascular, and rare diseases.

based on their structural characteristics, molecular size, and primary mechanisms of action, encompassing distinct groups such as monoclonal antibodies (mAbs), peptides, and other functional proteins, each tailored for specific therapeutic applications.

2.1.1. Monoclonal antibodies

Monoclonal antibodies are laboratory-produced proteins designed to bind with high specificity to particular antigens. Structurally, mAbs typically possess a Y-shaped structure with two heavy and two light chains, forming highly specific antigen-binding sites. Their mechanisms of action are diverse, including neutralizing pathogens or toxins, blocking receptor signaling, mediating antibody-dependent cellular cytotoxicity, or complem enting-dependent cytotoxicity, and acting as immune checkpoint inhibitors. mAbs are widely used in cancer immunotherapy, where they can act as checkpoint inhibitors (*e.g.*, pembrolizumab, nivolumab) to unleash the immune system's ability to attack tumor cells, leading to durable responses in various cancers (*e.g.*, melanoma, lung cancer)[24–27]. They are also extensively employed in treating autoimmune diseases (*e.g.*, adalimumab for rheumatoid arthritis, ustekinumab for psoriasis) by modulating specific immune responses[28,29], and in infectious diseases (*e.g.*,

Table 1 Representative biologics approved for clinical use.

Drug name	Drug type	Target/mechanism	Indication(s)	Approval year
Humira	Monoclonal antibody	TNF-α antagonist	Rheumatoid arthritis, Crohn's disease, etc.	2002
Adcetris	ADC	CD30-targeting ADC delivering MMAE toxin	Hodgkin lymphoma	2011
Spinraza	ASO	Modulates *SMN2* splicing	Spinal muscular atrophy (SMA)	2016
Exondys 51	ASO	Promotes exon 51 skipping in *DMD* gene	Duchenne muscular dystrophy (DMD)	2016
Hemlibra	Bispecific antibody	Bridges FIXa and FX to mimic FVIII function	Hemophilia A	2017
Onpattro	siRNA	Silences *TTR* mRNA expression	Hereditary transthyretin amyloidosis	2018
Zolgensma	Gene therapy (AAV)	AAV9 delivering functional *SMN1* gene	Spinal muscular atrophy (SMA)	2019
Enhertu	ADC	HER2-directed antibody−drug conjugate	$HER2^+$ breast and gastric cancer	2019
Inclisiran (Leqvio)	siRNA	Silences *PCSK9* mRNA	Hypercholesterolemia	2020
Comirnaty	mRNA	Encodes SARS-CoV-2 spike protein	COVID-19	2020
mRNA-1273	mRNA	Encodes stabilized SARS-CoV-2 spike protein	COVID-19	2020
Wegovy/Ozempic	Peptide	GLP-1 receptor agonist (semaglutide)	Obesity, type 2 diabetes	2021
Tezspire	Monoclonal antibody	Anti-TSLP, modulates airway inflammation	Severe asthma	2021
Mounjaro (Tirzepatide)	Peptide	Dual GIP and GLP-1 receptor agonist	Type 2 diabetes	2022
Qalsody (Tofersen)	ASO	Targets mutant *SOD1* mRNA	*SOD1*-related amyotrophic lateral sclerosis (ALS)	2023

palivizumab for respiratory syncytial virus[30]) or even migraine prevention (*e.g.*, erenumab[31]). Their high specificity translates to reduced off-target effects compared to traditional small molecule drugs.

2.1.2. *Peptides*

Peptides are oligomers composed of amino acids linked by covalent peptide bonds, typically ranging in length from 2 to 40 amino acid residues. Their specific three-dimensional conformations, dictated by linear or cyclic sequences, are fundamental to their diverse biological functions. Despite the relatively smaller molecular weight of some peptides, distinguishing them from traditional large, complex biologics (such as antibodies or recombinant proteins), they are widely regarded in drug discovery as a unique class of bioactive substances bridging small molecules and conventional biologics. This classification stems from their biological origin, precise sequence encoding, unique and intricate secondary/tertiary structures, and highly specific, high-affinity interaction modes with biological targets. Their distinct biological characteristics result in significant differences from traditional small molecule drugs in terms of mechanism of action and discovery strategies, aligning them more closely with the domain of biologics. Consequently, this review incorporates peptides into the scope of biologics for in-depth discussion, primarily considering the unique advantages and immense potential demonstrated by artificial intelligence technologies in peptide sequence design, conformational prediction, stability optimization, target identification, and assessments of drug-like properties. In comparison to full-length proteins, peptides typically exhibit a smaller molecular size, which often facilitates enhanced tissue penetration and simplifies their chemical synthesis. Their mechanisms of action are highly heterogeneous, encompassing roles as endogenous signaling molecules (*e.g.*, hormones, neurotransmitters, growth factors), direct antimicrobial agents, or modulators of protein-protein interactions *via* specific receptor or enzyme binding.

Peptide drugs are applied across a broad spectrum of therapeutic areas. In diabetes, GLP-1 receptor agonists like liraglutide and semaglutide have revolutionized treatment by enhancing glucose-dependent insulin secretion and promoting weight loss[32,33]. For osteoporosis, synthetic parathyroid hormone fragments such as teriparatide stimulate bone formation, offering an anabolic alternative to anti-resorptive agents[34]. In cancer, peptides like leuprolide (a GnRH agonist) are used to treat hormone-sensitive cancers such as prostate cancer[35], while novel peptide-drug conjugates are emerging for targeted chemotherapy[36]. Furthermore, peptides are being explored for their antimicrobial properties (*e.g.*, polymyxins[37]) and in various other conditions, including migraine prevention (*e.g.*, calcitonin gene-related peptide inhibitors like ubrogepant[38]). Their typically high specificity, relatively low immunogenicity, and favorable safety profiles make them attractive drug candidates, often with a more defined mechanism of action than small molecules.

2.1.3. *Functional proteins*

This broad category includes therapeutic proteins other than antibodies or short peptides, designed to mimic or replace naturally occurring proteins that are deficient, dysfunctional, or absent in disease. These recombinant proteins exhibit diverse structures, from complex enzymes to globular hormones and elongated growth factors, each engineered to perform specific biological

functions. Their mechanisms primarily involve the direct replacement of missing or non-functional endogenous proteins, catalysis of specific biochemical reactions, or modulation of cellular signaling pathways *via* receptor binding.

These recombinant proteins are crucial across a wide array of therapeutic areas. In metabolic and genetic disorders, they serve as life-saving replacement therapies. For instance, imiglucerase (recombinant glucocerebrosidase) is a prime example of enzyme replacement therapy for Gaucher disease, compensating for a genetic enzyme deficiency[39]. Similarly, recombinant human insulin (*e.g.*, insulin glargine) remains indispensable for managing diabetes, directly replacing the hormone lacking in patients[40]. Somatropin (recombinant human growth hormone) is vital for treating growth hormone deficiency and related conditions[41]. In hematological disorders, recombinant blood clotting factors like factor VIII (*e.g.*, octocog alfa) have transformed the management of hemophilia A, preventing or treating bleeding episodes[42]. Furthermore, hematopoietic growth factors like epoetin alfa (recombinant erythropoietin) stimulate erythrocyte production for anemia[43], while filgrastim (recombinant G-CSF) boosts neutrophil counts in chemotherapy-induced neutropenia[44]. Beyond replacement, some functional proteins act as immunomodulators, such as interferon-beta in the treatment of multiple sclerosis[45]. The key advantages of these drugs lie in their direct, specific action that addresses the underlying pathological cause, often leading to well-defined pharmacological effects and significant improvements in patient outcomes.

2.2. *Nucleic acid drugs*

Nucleic acid drugs harness nucleic acid molecules (DNA or RNA) as active pharmaceutical ingredients, directly modulating gene expression or protein function within biological systems. Based on their diverse mechanisms of action, these powerful therapeutics are broadly categorized into three main types: targeting nucleic acids for gene expression regulation, directly targeting proteins, and expressing therapeutic proteins *in situ*.

2.2.1. *Nucleic acids targeting gene expression regulation*

Nucleic acid drugs designed to target nucleic acids for gene expression regulation interact directly with other nucleic acid molecules, primarily messenger RNA or DNA, to control protein synthesis by either promoting or inhibiting translation. This category includes anti-sense oligonucleotides (ASO), which are synthetic single-stranded nucleic acids binding to specific mRNA sequences to block translation, induce mRNA degradation *via* RNase H, or modulate RNA splicing, thereby regulating protein levels. For instance, nusinersen (Spinraza) is a notable ASO approved for spinal muscular atrophy, modifying *SMN2* gene splicing to produce functional protein[46]. Small interfering RNA is another key component, being short, double-stranded RNA molecules that trigger RNA interference (RNAi), a natural cellular process. Once integrated into the RNA-induced silencing complex, siRNA directs sequence-specific cleavage of target mRNA, leading to post-transcriptional gene silencing. Patisiran (Onpattro), a siRNA drug, exemplifies this mechanism in treating hereditary transthyretin-mediated amyloidosis by silencing abnormal transthyretin production[47]. While still largely under investigation, microRNA (miRNA), typically endogenous regulators of gene expression, and small activating RNA (saRNA), which can paradoxically upregulate gene expression, also fall into this category. Furthermore, CRISPR/Cas systems represent programmable gene-editing tools that utilize guide RNA (gRNA) to direct a Cas enzyme (*e.g.*, Cas9) to specific DNA sequences for precise gene editing (insertions, deletions, or modifications), holding immense potential for correcting underlying genetic defects in diseases like sickle cell disease and beta-thalassemia, with numerous clinical trials underway[48].

2.2.2. *Nucleic acids directly targeting proteins*

The second category of nucleic acid drugs consists of those that directly target proteins, exerting their therapeutic effects by binding to and modulating the function of specific protein targets rather than primarily regulating gene expression. The prime example here is aptamers, which are single-stranded DNA or RNA oligonucleotides designed to fold into unique three-dimensional structures. These structures enable them to bind to specific protein targets with high affinity and selectivity, functionally analogous to antibodies. Their binding can inhibit protein function or block protein−ligand interactions. An early clinical success, pegaptanib (Macugen), an aptamer approved for wet age-related macular degeneration, directly binds to and inhibits vascular endothelial growth factor, preventing pathological neovascularization[49]. Aptamers are also being explored for diverse applications in oncology, cardiovascular diseases, and infectious diseases.

2.2.3. *Nucleic acids for protein expression*

Finally, the third category encompasses nucleic acid drugs designed to express therapeutic proteins within the body, effectively transforming the patient's own cells into drug factories. The most impactful example of this mechanism is *in vitro* transcribed mRNA. These synthetic mRNA molecules, often chemically modified for enhanced stability and translational efficiency, are introduced into host cells. Once internalized, they hijack the cellular ribosomal machinery to translate the encoded genetic information into a desired therapeutic protein. The most prominent application of mRNA has been in vaccinology, particularly the COVID-19 mRNA vaccines (*e.g.*, Moderna's mRNA-1273 and Pfizer/BioNTech's BNT162b2[50,51]), which instruct host cells to produce the SARS-CoV-2 spike protein, eliciting a robust and protective immune response. Beyond infectious disease prophylaxis, mRNA therapeutics are also actively being developed for cancer immunotherapies, protein replacement therapies for genetic disorders, and as a delivery mechanism for *in vivo* gene-editing enzymes, underscoring their vast therapeutic potential.

2.3. *Delivery systems of macromolecules*

Efficient *in vivo* delivery of biological macromolecules is a complex, multi-step process that requires ensuring activity preservation, precise targeted delivery, effective cellular internalization, and appropriate intracellular trafficking and release. To overcome the inherent challenges of macromolecule drugs, especially for nucleic acid drugs, including their instability, degradability, low permeability, and lack of intrinsic targeting, current *in vivo* delivery systems can be primarily categorized into two main strategies: carrier-based and conjugate-based.

2.3.1. *Carrier-based delivery system*

Carrier-based delivery systems encapsulate macromolecule drugs within independent carriers to protect them from degradation, enhance stability, and facilitate passage across biological barriers while enabling targeted delivery. This category encompasses diverse and evolving systems. Lipid nanoparticles (LNPs), for

instance, are self-assembled nano-carriers highly effective for nucleic acid drug delivery (*e.g.*, mRNA, siRNA), famously serving as the crucial platform for COVID-19 mRNA vaccines, demonstrating unprecedented efficacy[50,51]. Viral vectors leverage genetically modified, non-pathogenic viruses to efficiently transduce target genes into specific cells. Adeno-associated viruses (AAV) have emerged as dominant vectors in gene therapy, with approved drugs like Zolgensma for spinal muscular atrophy and Luxturna for inherited retinal disease[52,53]. Lentiviruses (LVs), which integrate into the host genome for long-term stable gene expression, underpin *ex vivo* gene therapies such as Kymriah and Yescarta for certain hematological malignancies[54,55]. While adenoviruses (ADVs) are also utilized for their high transduction efficiency and broad cellular tropism. More recently, extracellular vesicles (EVs), naturally secreted nanoscale membrane-bound vesicles, are gaining significant attention. These endogenous drug delivery systems, including exosomes (30−150 nm), microparticles/microvesicles (100−1000 nm), apoptotic bodies/blebs (up to 1−5 μm), and large oncosomes, show promising potential for delivering nucleic acids or proteins to specific tissues in areas like cancer therapy and regenerative medicine, although largely still in preclinical and early clinical development.

2.3.2. *Conjugate-based delivery system*

Conjugate-based targeted delivery directly links a targeting molecule to a biologic drug *via* a chemical bond, creating an innovative drug with inherent targeting capabilities. This integrated design aims for precise recognition and action on specific cells or tissues *in vivo*. Antibody−drug conjugates (ADC) are prime examples, which pair a highly cytotoxic small molecule with a monoclonal antibody *via* a chemical linker. The antibody precisely identifies and targets specific antigens on tumor cell surfaces, delivering the drug directly into cancer cells while minimizing harm to healthy tissues. Kadcyla (trastuzumab emtansine), used for HER2-positive breast cancer, exemplifies this strategy[56].

Peptide−drug conjugates (PDC) involve conjugating a bioactive peptide, often chosen for its targeting or cell-penetrating abilities—with a drug molecule. These systems exploit the peptide's specific binding or cellular permeation properties for effective delivery. Lutathera (lutetium (^{177}Lu) oxodotreotide), a notable PDC for neuroendocrine tumors, links a somatostatin analog peptide to a radioisotope, allowing targeted radiation to tumor cells expressing somatostatin receptors[57]. Increasingly, nucleic acid conjugates are gaining attention. These involve coupling nucleic acid drugs (like siRNA or ASO) with specific targeting ligands, such as GalNAc, peptides, or antibody fragments. This enhances the nucleic acid drugs' *in vivo* stability and targeting specificity, significantly improving its efficiency in reaching particular cells or tissues. Inclisiran (Leqvio) is a leading example: this GalNAc-conjugated siRNA efficiently delivers the siRNA to hepatocytes for hypercholesterolemia, where it silences *PCSK9* mRNA, resulting in reduced LDL-C levels[58].

In recent years, AI has rapidly transformed the landscape of drug discovery, with multiple AI-designed small-molecule therapeutics advancing into human studies. Notable examples include DSP-1181 (Exscientia) for obsessive-compulsive disorder[59]; EXS4318 (Exscientia/Bristol-Myers Squibb) for immune and inflammatory diseases[60]; DSP-0038 (Exscientia) for Alzheimer's-related psychosis[61]; EXS21546 (Exscientia/Evotec) targeting the A2a receptor in oncology[62]; and Rentosertib (ISM001-055), a TNIK inhibitor developed by Insilico Medicine for idiopathic pulmonary fibrosis[63]. Rentosertib has already demonstrated promising results in a phase IIa clinical trial, marking a milestone in the clinical application of AI and providing one of the clearest validations to date that AI-designed molecules can advance successfully through clinical development[64].

By contrast, the development of biologics and their delivery systems, including proteins, peptides, nucleic acids, and carriers such as lipid nanoparticles, AAV, and ADC, remains more challenging due to complex molecular structures, difficulties in ensuring stability and targeted delivery, and high attrition rates in clinical trials. AI-driven biologics pipelines are therefore still at an early stage. Several AI-designed candidates have already advanced into clinical evaluation, spanning diverse therapeutic modalities. Notable examples include Evaxion's personalized peptide-based cancer immunotherapy for metastatic melanoma (phase II)[65], ZielBio's plectin-targeting antibody for solid tumors (phase II)[66], and PharmCADD's mRNA vaccine candidate against SARS-CoV-2 (phase II)[67]. Early-stage programs are also emerging, such as Peptilogics' peptide antibiotic for periprosthetic joint infection[68], SparX Therapeutics' monoclonal antibody targeting Claudin 18.2 for gastric cancer[69], and an additional mRNA vaccine candidate from PharmCADD[70], all of which are currently in phase I trials. These emerging examples illustrate that, while still nascent, AI-guided biologic drug discovery is beginning to achieve tangible clinical progress.

The introduction of AI offers promising solutions to the unique challenges in this space: deep learning models can predict and optimize macromolecular structures, guide the rational design of delivery systems, and assist in clinical trial planning by leveraging diverse datasets to improve predictive accuracy. These advances are beginning to reshape the development paradigm for biologics, providing a bridge from traditional experimental approaches to data-driven, AI-guided strategies. Building on this foundation, the following section reviews recent studies that exemplify AI-driven approaches in biologic drugs and delivery system design.

3. Recent major AI technologies evolutions in biologic drugs

The concurrent explosion in biological data and computational power has propelled a groundbreaking transformation in biologic drug discovery, with AI synergizing with bioinformatics. AI is no longer merely an accelerator for existing workflows but a transformative force catalyzing entirely new paradigms for designing and engineering biologics. By discerning complex patterns from massive datasets, AI is fundamentally reshaping our ability to understand, predict, and create biologic therapeutics. This section delineates the pivotal breakthroughs of prominent AI methodologies, charting their progressive evolution from sequence understanding to structure prediction, and ultimately to the frontier of functional *de novo* design and engineering.

3.1. *Biological sequence language models*

The application of natural language processing (NLP)[71] to molecular biology represents a conceptual leap, treating the constituent monomers of macromolecules—20 canonical amino acids for

proteins and 4 primary nucleotides for nucleic acids—as a vocabulary. This paradigm allows deep learning models to learn the complex grammar governing biological sequences, moving beyond superficial statistical analysis to a deep, contextual understanding. This approach has given rise to a powerful class of foundational models known as biological language models[72].

Initially, these models, architecturally based on the Transformer (Fig. 2A), were trained on vast, unlabeled publicly accessible sequence databases such as UniProt[73] (ftp.uniprot.org/pub/databases/uniprot/) for proteins and GenBank[74] (ftp.ncbi.nlm.nih.gov/genbank/) for nucleic acids through self-supervised learning. The primary training objective, masked language modeling (MLM)[75], compels the model to predict masked-out tokens (amino acids or nucleotides) based on their surrounding context. This process forces the model to internalize profound knowledge regarding evolutionary conservation, structural constraints, and biophysical plausibility. The output is a high-dimensional vector representation for each residue, which captures latent biological information and significantly enhances performance on diverse downstream tasks[76,77].

In protein science, the ESM (evolutionary scale modeling) series became a cornerstone (Fig. 2B), with models like ESM2[78] demonstrating an exceptional ability to predict the functional effects of mutations based on a sequence's naturalness or pseudo-log-likelihood (pLL) score, providing a high-throughput computational filter for protein engineering[79]. Generative biological language models, represented by the ProGen series[80–83], exemplify the next logical step: a transition from comprehension to generation. These methods primarily utilize a Transformer decoder architecture and are trained *via* a next-token prediction task, generating sequences amino acid by amino acid based on the preceding fragment. By conditioning the generation process on functional annotations (*e.g.*, enzyme commission numbers), ProGen can create novel, biologically plausible protein sequences with specific, desired functions[81]. Meanwhile, researchers have also tailored Transformer architectures for the precise prediction of complex molecular interactions; for instance, the physics-inspired sliding Transformer (PISTE)[23] model exemplifies this by introducing a physics-inspired attention mechanism to predict the binding of the critical TCR–pMHC ternary complex.

The recent ESM3[84] represents a significant advance into multimodality[85–87], integrating 3D structural information directly into its training[88], enabling it to generate novel, functional proteins like esmGFP through complex, structure-aware prompts. In parallel, the field of genomics has seen the development of models like DNABERT[89–92], which learned the fundamental syntax of DNA from the human genome. More advanced genomic foundation models like Evo[93], trained on hundreds of billions of nucleotides, leverage long-context architectures to understand the intricate, long-distance regulatory logic governing entire genomic regions, enabling zero-shot function prediction across DNA, RNA, and proteins from a single model.

However, these models are not without limitations. Their predictive accuracy is intrinsically tied to the quality and diversity of the training data; for instance, models trained on UniProt may struggle to accurately represent proteins from sparsely sampled regions of the evolutionary tree. Furthermore, while their zero-shot capabilities are powerful for general properties like stability, they often fall short in predicting highly specific functions that are under-represented in the pre-training corpus. This gap has motivated a shift from purely unsupervised models to those that can be efficiently fine-tuned on smaller, task-specific labeled datasets, a central theme in the engineering applications discussed in Section 5. The immense computational cost of training and running these

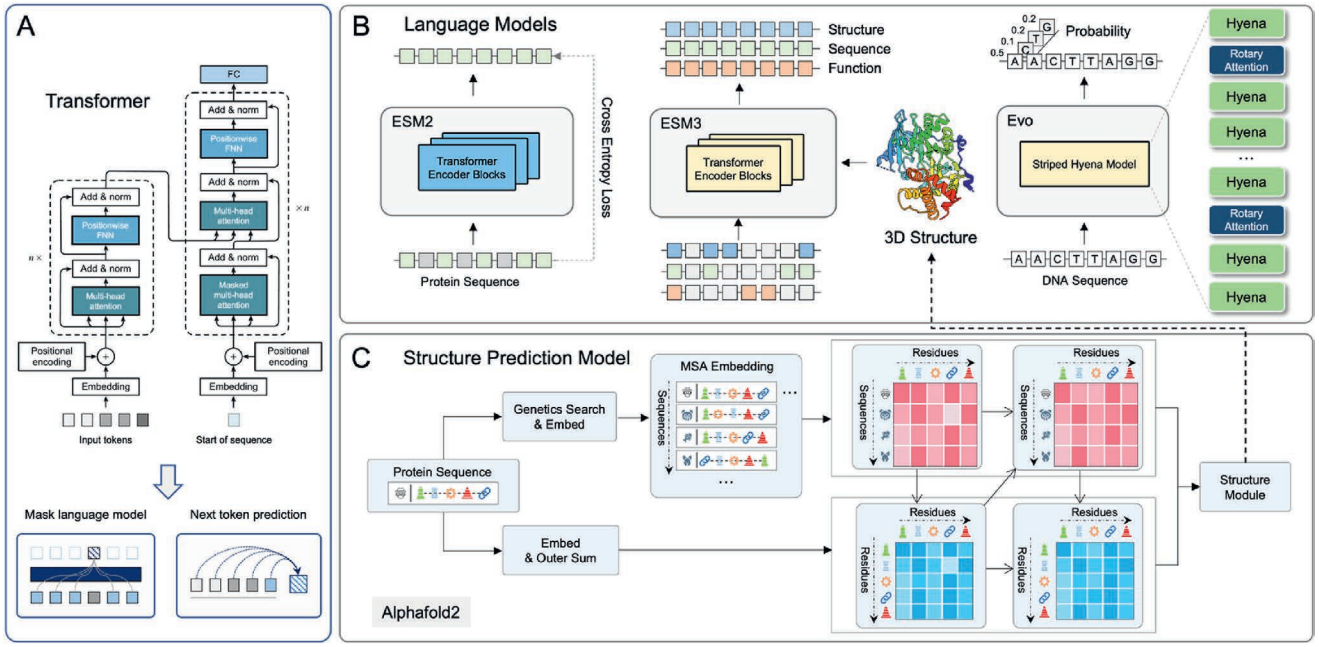

Figure 2 Algorithm development and application for sequence and structural biological data. (A) The Transformer architecture serves as the foundational backbone, utilizing attention mechanisms to decipher biological sequence patterns. (B) Biological language models have advanced from analyzing protein sequences (ESM2) to integrating multimodal data (ESM3) and genomic contexts (Evo). (C) Structure prediction models like AlphaFold2 leverage evolutionary information from MSAs to accurately generate 3D protein conformations.

large-scale models also presents a practical barrier, driving research into more efficient architectures and knowledge distillation techniques.

3.2. From sequence to structure

While language models decode one-dimensional sequence information, biological function is ultimately dictated by three-dimensional structure[94–96]. For decades, accurately predicting a macromolecule's 3D fold from its primary sequence stood as a grand challenge in biology[97]. The breakthrough arrived with end-to-end deep learning models, which transformed structural biology from a predominantly experimental science to one where high-accuracy prediction is computationally accessible (Fig. 2C).

The paradigm shift was spearheaded by DeepMind's AlphaFold2[98] and the Baker laboratory's RoseTTAFold[99]. These models ingeniously fuse co-evolutionary information, derived from multiple sequence alignments (MSA), with the physical and geometric constraints of molecular structures. Their dual-track Evoformer architecture employs attention mechanisms to simultaneously reason over 1D sequence relationships and a 2D map of inter-residue distances, iteratively refining a 3D structural hypothesis with atomic-level accuracy. This revolution quickly expanded beyond single proteins. Inspired by these principles, models like RhoFold[90,100] were developed to tackle the complex, non-canonical folds of RNA. The current frontier, exemplified by AlphaFold3[101] and RoseTTAFold All-Atom[102], has unified these domains, enabling the prediction of vast multi-component complexes involving proteins, DNA, RNA, small-molecule ligands, and ions, thus providing an unprecedented atomic-level view of cellular machinery.

However, a critical limitation of these landmark models is that they typically predict a single, static, low-energy conformation. This overlooks the fact that proteins and nucleic acids are dynamic machines that flex, bend, and shift to execute their functions. Capturing this conformational ensemble is the next frontier. Pioneering models like BioEMU[103] are designed to address this by learning a distribution of conformations from a combination of static structures, experimental ensemble data, and, crucially, molecular dynamics (MD) simulation trajectories. This allows for the modeling of processes invisible to static models, such as allosteric regulation and ligand binding pathways.

The impact of these predictive tools extends far beyond basic biological inquiry. Within the drug development workflow, they serve two primary roles. First, they act as a crucial *in silico* validation filter. As will be detailed in Section 4, a computationally designed molecule can be fed into a model like AlphaFold2; a high-confidence prediction that matches the intended design provides strong evidence that the molecule is physically plausible, thereby de-risking costly experimental validation. Second, they provide the essential structural scaffolds required for structure-based engineering and *de novo* design. Despite their power, their accuracy is heavily dependent on the availability of deep MSA, posing a challenge for orphan proteins or highly novel designs with no evolutionary history. Furthermore, the prediction of dynamic states and the precise modeling of non-proteinaceous components, while rapidly improving, remain active areas of research with higher uncertainty than that for well-conserved, single-chain proteins.

3.3. Generative models for de novo biologics molecular design

Leveraging the deep understanding of sequence and structure, the ultimate goal of AI is to create entirely novel biologic molecules with bespoke functions from first principles[104,105]. This marks a shift from analyzing existing biology to generating new biology. Generative models, particularly diffusion and autoregressive models, are at the forefront of this creative endeavor[106].

Diffusion models[107,108], adapted from image generation, operate by learning to reverse a noise-adding process. They start with a random cloud of points and progressively denoise it into a coherent, physically plausible molecular backbone. This process is highly controllable, allowing designers to impose geometric constraints, such as the shape of a binding pocket or the symmetry of an oligomer. RFdiffusion[109] is a paradigmatic example in protein design, capable of generating a vast array of novel backbones. Once a backbone is created, the workflow proceeds to the inverse folding problem: designing a sequence that will adopt that specific structure[110–115]. Here, graph neural network-based models like ProteinMPNN[116] have proven exceptionally potent, painting a compatible amino acid sequence onto the target backbone with high accuracy. Alternative paradigms that unify these steps are also emerging. Chroma[117], for instance, co-designs sequence and structure in a single, unified diffusion process, while methods like relaxed sequence optimization (RSO)[118] use the backpropagated gradients from a structure predictor like AlphaFold2 to directly optimize a sequence towards a target structure. To further enhance the sophistication of generative design in the protein domain, recent latent diffusion frameworks such as DiMA[119] now operate on the learned continuous representations of protein language models, enabling powerful and controllable sequence generation for complex tasks like fold-conditioned design.

While the protein domain has historically been at the forefront of these generative methods, the field of nucleic acid design is now rapidly advancing with the adoption of similarly sophisticated techniques. Moving beyond initial applications, a new generation of models is tackling the direct *de novo* design of 3D RNA structures and their corresponding sequences. For instance, RNA-Frameflow[120] adapts flow-matching techniques to unconditionally generate novel RNA backbones, demonstrating the power of these methods for exploring new structural space. Building on this, models like RiboFlow[121] employ a synergistic flow-matching approach for the conditional co-design of both RNA sequence and structure, aiming to create functional molecules that bind specific ligands. In parallel, diffusion models are being tailored to the unique geometry of nucleic acids; RIdiffusion[122], for example, utilizes a hyperbolic discrete diffusion process for the challenging task of RNA inverse folding. These cutting-edge methods signify a pivotal shift, moving nucleic acid design from sequence-centric approaches to a new era of structure-aware, generative 3D modeling.

The power of generative models lies in their ability to explore a design space far vaster than that accessible to nature or manual engineering. However, the "*in silico* to *in vivo*" gap remains the

most significant challenge. A generated protein may fold correctly according to AlphaFold but may not express well in cells, be soluble, or could be immunogenic. The current generation of models is trained primarily on structural data and lacks an intrinsic understanding of these complex biological properties. Therefore, the end-to-end *de novo* design workflow is an iterative, multi-stage process: (1) Constrained generation of a backbone (*e.g.*, RFdiffusion), followed by (2) sequence design (*e.g.*, ProteinMPNN), which is then subjected to (3) *in silico* validation (*e.g.*, AlphaFold2). The most promising candidates from this computational pipeline then enter the experimental validation bottleneck, where their real-world properties are tested, and the data is used to inform the next design cycle, as discussed in Section 5. Effective drug design strategies increasingly rely on the synergistic integration of these computational tools to maximize experimental success rates.

3.4. AI methodologies for delivery system

Beyond the design of the therapeutic macromolecule itself, its effective and safe delivery presents a distinct and equally critical challenge. Artificial intelligence is also revolutionizing this space, particularly in the design of non-viral vectors like LNPs. As LNPs are complex formulations of small molecules, the AI methodologies employed differ from those used for macromolecules, focusing on learning the intricate, non-linear relationships between chemical structure, formulation parameters, and biological function.

The primary application of AI in this domain involves building discriminative (or predictive) models to navigate the vast chemical space of potential LNPs components, most notably the ionizable lipids that are critical for efficacy. These models are trained to predict key delivery-related properties, such as transfection efficiency or *in vivo* biodistribution, directly from molecular structure. To achieve this, lipid structures are typically encoded into machine-readable formats, such as molecular fingerprints (*e.g.*, extended connectivity fingerprints, ECFPs) or graph representations. These encoded data are then used to train a wide range of models, from classic machine learning algorithms like gradient boosting frameworks to more complex deep learning architectures like directed message passing neural networks (DMPNN). The core function of these models is to act as a high-throughput computational screening tool, enabling the rapid identification of promising lipid candidates from extensive virtual libraries, thereby prioritizing experimental resources for the most viable options.

Building upon this predictive foundation, the field is advancing towards generative models for the *de novo* design of novel delivery components. In contrast to screening pre-existing or enumerated molecules, generative approaches aim to create entirely new ionizable lipids that are intrinsically optimized for desired properties. A key innovation in this area is the incorporation of synthetic accessibility as a direct objective in the design process. Models based on reinforcement learning, Monte Carlo tree search (MCTS), or deep generative frameworks are trained on libraries of known chemical building blocks and reaction rules. This allows them to assemble novel, high-performing lipid structures that are not only predicted to be effective but are also composed of readily available precursors and connected *via* high-yield chemical reactions, thus bridging the gap between computational design and practical laboratory synthesis. These generative and predictive models often work in tandem, forming a powerful design loop where novel molecules are generated and then immediately scored for efficacy, stability, and synthesizability.

4. AI-driven *de novo* design of biologic therapeutics

In the preceding chapters, we systematically traced the methodological evolution of artificial intelligence in biological macromolecular science. We charted a course from the initial breakthroughs in deciphering sequence language and predicting three-dimensional structures to the ultimate frontier of creating functional molecules. We now pivot from this theoretical foundation to the practical impact of these technologies (Fig. 3A). This section aims to illuminate the real-world applications and milestone achievements of AI in specific domains of biologic drug design. We will focus on the pivotal role of AI in key areas such as *de novo* enzyme design, antibody engineering, therapeutic peptide development, and the precision engineering of nucleic acid therapies (Fig. 3B). We will first outline the modern, generalized workflow that underpins these efforts before delving into specific domains.

4.1. The computational paradigm of de novo biologic design

The transformation from computational hypothesis to therapeutic candidate represents a fundamental shift in biologic drug discovery, enabled by the convergence of generative artificial intelligence and structural biology. Modern *de novo* design workflows for both protein and nucleic acid therapeutics follow a unified computational paradigm that integrates three core methodological components.

The foundational layer employs generative models for molecular backbone creation. For proteins, diffusion-based models such as RFdiffusion[109] generate novel, foldable backbones through progressive denoising of random atomic coordinates, with the process controllable through geometric constraints specific to desired functional sites. Similarly, for nucleic acids, models like RNA-Frameflow[120] and RiboFlow[121] apply flow-matching techniques to generate three-dimensional RNA structures. These generative processes operate in continuous coordinate space, enabling precise control over structural features while exploring regions of molecular space inaccessible to natural evolution.

The second component addresses the inverse folding problem through sequence design. Graph neural network architectures, exemplified by ProteinMPNN[116] for proteins and RhoDesign[123] for nucleic acids, learn the complex mapping between three-dimensional structure and compatible sequences. These models process molecular backbones as geometric graphs, predicting amino acid or nucleotide identities that satisfy both local chemical environments and global folding stability requirements.

The critical validation layer employs high-accuracy structure predictors as computational filters. AlphaFold2[98] and its derivatives provide atomic-level validation for protein designs, while specialized models assess nucleic acid structures. These predictors generate confidence metrics that serve as powerful selection criteria, dramatically reducing the experimental validation burden by eliminating designs unlikely to adopt their intended conformations. This iterative computational pipeline, transitioning from generation through sequence design to validation, establishes a robust framework for creating novel biologics with predetermined structural and functional properties.

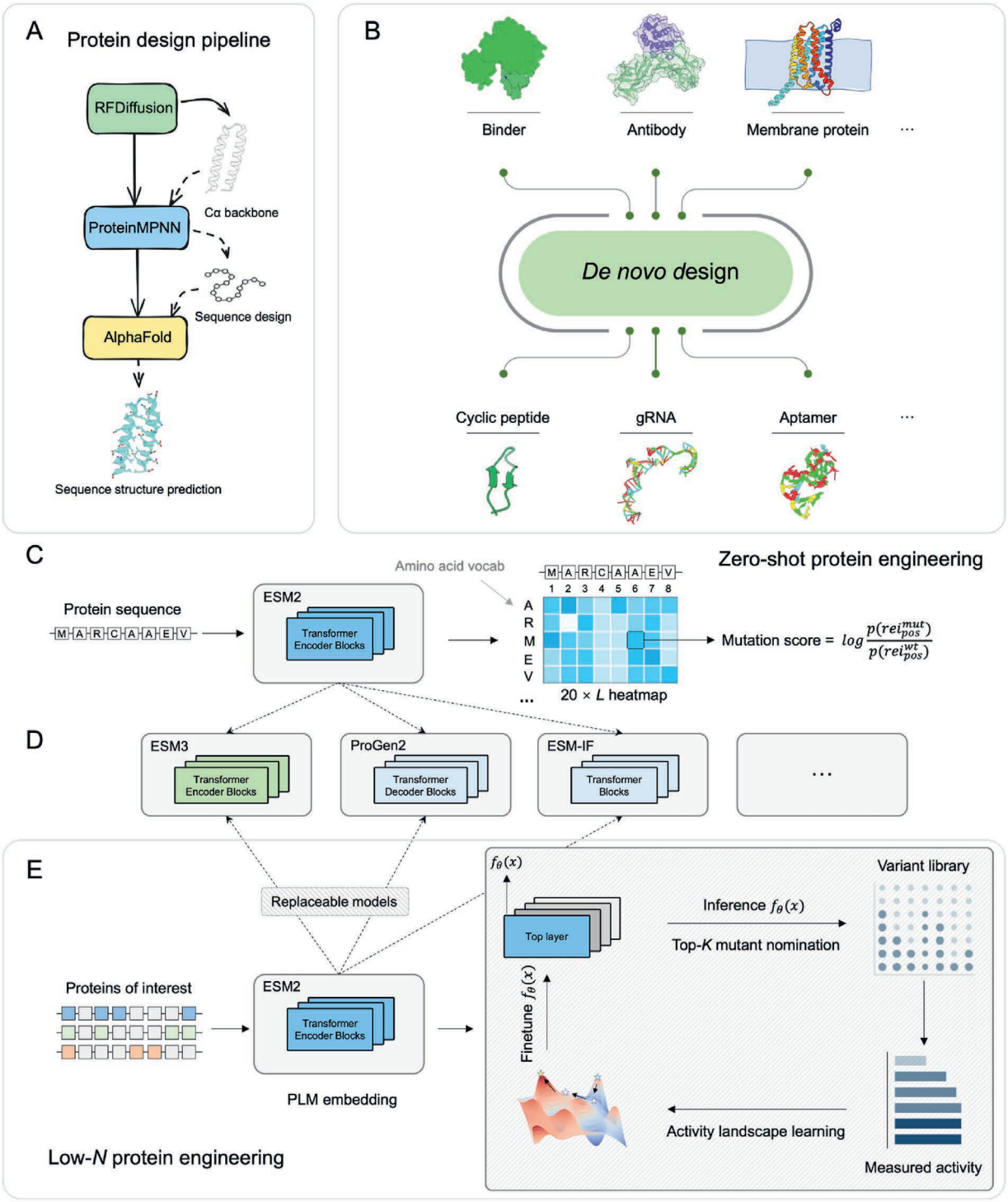

Figure 3 Algorithms and applications of *de novo* protein design and engineering. (A) *De novo* biologic drug design pipeline and (B) main therapeutic application scenarios. (C, D) Zero-shot protein engineering workflow and interchangeable protein base models, and (E) low-N lightweight supervised fine-tuning directed evolution methods.

4.2. Designing protein therapeutics

4.2.1. Designing high-affinity binders and membrane protein modulators

The design of proteins that bind specific molecular targets with high affinity and selectivity represents a cornerstone application of AI-driven protein engineering[124–126]. Recent advances have dramatically improved success rates through the integration of deep learning models into established design pipelines. The Baker group[127] demonstrated that incorporating AlphaFold2 or RoseTTAFold for assessing folding probability and binding configuration increases experimental success rates by nearly tenfold compared to traditional physics-based methods. This improvement stems from the models' ability to provide reliable

confidence metrics, particularly predicted local distance difference test (pLDDT) and predicted aligned error (pAE) scores, which effectively filter designs before experimental validation.

The application to therapeutic targets has yielded notable successes. Using RFdiffusion with precise geometric constraints, researchers designed proteins that specifically bind and neutralize multiple snake venom three-finger toxin subclasses[128]. The computational generation of binding backbones complementary to the β-sheet structures of target toxins, followed by AlphaFold2-based filtering, enabled the creation of broadly neutralizing antitoxins. This approach demonstrates the capacity to address previously intractable targets through computational design alone. Complementing diffusion-based approaches, BindCraft[129] leverages trained structure prediction networks for the one-shot design of high-affinity binders, successfully targeting challenging antigens such as allergens, cell-surface receptors, and multi-domain nucleases with nanomolar affinity.

Membrane proteins present unique challenges due to their hydrophobic nature and requirement for lipid bilayer environments[130,131]. The design of GPCR modulators exemplifies the sophistication of current methods[132,133]. By combining motif-guided RFdiffusion with high-throughput screening platforms, researchers successfully created small-protein agonists and antagonists for therapeutically relevant GPCRs including CXCR4 and GLP1R[134]. Cryo-EM validation confirmed that these designed proteins precisely engage their intended receptor conformations. Alternative approaches leverage protein language models that are less dependent on structural templates. Methods based on backpropagation, such as RSO, have also demonstrated the capability in designing membrane proteins with complex topologies. By coupling AlphaFold2's predictive power with a differentiable sequence space, researchers[135] successfully transplanted the functional domains of membrane proteins onto soluble scaffolds, achieving nanomolar binding affinities without requiring physical constraints or experimental optimization. The MeMDLM[136] masked diffusion language model learns sequence features of natural membrane proteins[137] to directly generate novel sequences with correct transmembrane topology, providing a complementary strategy when structural data is limited.

4.2.2. *Designing therapeutic antibodies*

The *de novo* design of antibodies requires addressing the unique architectural complexity of these molecules, particularly the hypervariable complementarity-determining regions that dictate specificity. Current AI approaches bifurcate into sequence-based and structure-based methodologies, each offering distinct advantages.

Sequence-based models leverage the vast natural antibody repertoire to learn distributional properties, enabling novel sequence generation. PALM-H3[138], a specialized language model for complementarity-determining regions (CDR), especially for CDRH3 generation, demonstrates the power of this approach. After pre-training on massive antibody sequence corpora, fine-tuning for specific antigens enables generation of CDRH3s with sequences divergent from natural antibodies yet exhibiting superior affinity and neutralization activity. The AbBFN2[139] framework extends this capability through Bayesian flow networks that concurrently process and generate antibody sequences alongside metadata, including species origin, germline lineage, and developability scores. This multimodal capability enables complex operations including sequence inpainting, humanization, and multi-objective optimization within a unified framework.

Structure-based approaches achieve atomic-level precision for epitope-specific binding. Bennett[140] and colleagues fine-tuned RFdiffusion to enable complete *de novo* design of antibody variable regions targeting specific antigen epitopes. Requiring only antigen structure and epitope information as input, this method generates entirely novel antibodies with correct immunoglobulin folds and precise binding modes. Cryo-EM validation revealed high concordance between designed antibodies and computational models at atomic resolution. The IgGM[141] framework advances this through co-design of antibody−antigen complexes, simultaneously generating both sequence and structure while directly predicting binding conformation. This integrated approach enables flexible application to various design scenarios, from complete CDR generation to targeted CDRH3 optimization.

4.2.3. *Computational design of therapeutic peptides*

Therapeutic peptides occupy a unique chemical space between small molecules and large proteins, offering high target specificity with potentially favorable pharmacological properties. AI-driven peptide design strategies have evolved to address the dual challenges of achieving high binding affinity while maintaining drug-like properties including stability and membrane permeability[142,143].

Sequence-based approaches treat peptides as biological sentences, leveraging language models to generate novel sequences with desired properties. The PepPrCLIP[144] workflow exemplifies this strategy through a two-step process. First, Gaussian perturbations applied to the latent space of protein language models generate large libraries of natural-like peptide sequences. Subsequently, a discriminative model inspired by contrastive language-image pre-training screens these candidates for high and selective binding activity. Operating entirely on sequence information, this method has successfully designed linear peptides that inhibit disease-driving targets with diverse conformational states, including intrinsically disordered proteins traditionally considered undruggable.

Specialized frameworks address specific therapeutic applications. For antimicrobial peptides, GPT-based architectures[145] integrate prompt tuning with reinforcement learning to enable multi-parameter optimization. This allows simultaneous optimization of antimicrobial activity, hemolytic toxicity, and plasma stability, generating peptides that are potent, minimally toxic, and resistance-resistant. Beyond antimicrobial applications, AI is also advancing the design of peptide vaccines. For example, Feng et al.[146] employed deep learning to design a multi-valent peptide vaccine against SARS-CoV-2, demonstrating broad immune protection. The PepINVENT[147] model expands chemical diversity by seamlessly incorporating non-natural amino acids through bespoke chemical language representations, vastly enriching the functional potential of designed peptides.

Structure-based methods provide atomic-level design precision for high-affinity binding. Utilizing RFdiffusion with strategies including partial diffusion and unconstrained hallucination, researchers designed protein scaffolds binding bioactive helical peptides with picomolar affinity[148]. For cyclic peptides, which offer exceptional stability but present design challenges, the AfCycDesign[142] workflow repurposes AlphaFold2 by modifying its positional encoding matrix to introduce cyclization constraints. This enabled the hallucination of over ten thousand novel cyclic peptide backbones with experimentally confirmed atomic accuracy.

4.2.4. Therapeutic enzymes

The design of therapeutic enzymes demands the realization of complex catalytic functions beyond simple binding, requiring precise substrate recognition, stable active site conformations, and efficient catalytic cycles. This represents one of the most challenging applications of AI-driven protein engineering[149,150].

Recent successes demonstrate the potential for creating novel therapeutic functionalities. Computational design of proteins that specifically bind and clear small-molecule drugs exemplifies this capability[151]. Using the COMBS algorithm for initial scaffold construction followed by strategies including backbone cyclization and D-amino acid synthesis, researchers created proteins that specifically bind anticoagulants and anticancer drugs. These designed proteins, particularly the D-amino acid-configured variants, demonstrated remarkable drug clearance efficiency in mouse models while remaining non-immunogenic.

Language models provide a powerful engine for creating entirely new gene editors. By training on the most comprehensive CRISPR-Cas atlas assembled from metagenomic datasets, researchers developed models capable of generating functional, novel CRISPR-Cas proteins[152]. The generated OpenCRISPR-1 editor exhibited profound sequence divergence from known proteins yet demonstrated precise gene-editing activity in human cells. The CasGen[153] model advances this through margin-based latent space regularization and Bayesian optimization, ensuring generated sequences reside within high-probability functional regions while maintaining structural validity through AlphaFold2 validation.

Beyond direct enzyme design, AI enables the enhancement of existing biological systems through regulatory component engineering. The design of MLH1-SB, a small protein inhibitor of the mismatch repair pathway, demonstrates this systems biology approach[154]. Using RFdiffusion to design proteins that specifically bind the MLH1-PMS2 heterodimer interface, researchers created inhibitors that shield prime editing from MMR interference, increasing editing efficiency by several-fold to over an order of magnitude. This work showcases the potential for AI-designed modulators to optimize complex intracellular enzymatic networks.

4.2.5. Current achievements and challenges in protein therapeutics

The integration of AI into protein therapeutic design has yielded transformative achievements. Success rates for *de novo* protein design have increased dramatically, with computational filtering reducing experimental validation requirements by orders of magnitude. The ability to design proteins targeting previously intractable molecules, including membrane proteins and intrinsically disordered regions, has expanded the druggable proteome. Multi-objective optimization frameworks now enable simultaneous engineering of binding affinity, stability, and developability properties, addressing downstream manufacturing challenges during the design phase.

However, significant challenges constrain the full realization of AI's potential in protein therapeutics. The prediction of complex biological properties remains limited, with reliable models for immunogenicity, *in vivo* stability, and pharmacokinetics still in development. The vast majority of successful designs still require extensive experimental optimization to achieve clinical viability. The reality gap between *in silico* predictions and biological function persists, particularly for properties emerging from protein−protein interactions or cellular context.

The core challenge lies in the multi-scale nature of protein function. While AI excels at predicting local structural features and binding interfaces, capturing emergent properties arising from dynamics, allostery, and cellular environment remains elusive. Current models trained primarily on static structural data lack an intrinsic understanding of temporal dynamics critical for enzymatic function. Furthermore, the optimization of multiple conflicting objectives, such as simultaneously maximizing binding affinity while minimizing aggregation propensity, requires navigation of complex fitness landscapes where improvements in one property often compromise others. Addressing these challenges will require integration of dynamics-aware models, improved representation of cellular contexts, and development of more sophisticated multi-objective optimization algorithms that can effectively explore high-dimensional Pareto fronts.

4.3. Nucleic acid therapeutics

The design principles for nucleic acid therapeutics have traditionally been governed by the foundational rules of Watson-Crick base pairing and empirical screening methodologies. For instance, the design of RNAi agents like ASO and siRNA primarily relied on sequence complementarity to the target mRNA, often overlooking the complex three-dimensional structures that dictate target accessibility and lead to high failure rates[155,156]. Similarly, the discovery of nucleic acid aptamers has been dominated by the laborious, time-consuming, and costly systematic evolution of ligands by exponential enrichment (SELEX) process, which explores only a minute fraction of the vast theoretical sequence space[157]. The advent of generative artificial intelligence is catalyzing a paradigm shift, moving beyond these limitations to enable the *de novo* design of nucleic acid drugs with novel sequences, structures, and enhanced functionalities.

4.3.1. mRNA therapeutics

The therapeutic potential of mRNA, prominently demonstrated by COVID-19 vaccines, depends critically on optimized sequence design encompassing both coding and regulatory regions. AI-driven approaches have transformed mRNA design from empirical selection to systematic engineering of sequences with enhanced stability and expression[158,159].

Generative models based on Transformer architectures treat mRNA sequence generation as a language modeling task. Models including GEMORNA[158] and GenerRNA[160], pre-trained on massive RNA sequence datasets, learn the intrinsic grammar of RNA encompassing sequence motifs, structural propensities, and regulatory codes. These models generate novel coding sequences and untranslated regions that go beyond simple codon optimization to learn complex, context-dependent patterns enhancing both translation efficiency and stability. Specialized models like Smart5UTR[159] and platforms developed by Morrow and colleagues[161] focus on generating novel UTRs, demonstrating significant enhancements in protein expression and mRNA durability. This holistic approach enables co-design of all mRNA components as an integrated system, optimizing the interplay between sequence, structure, and function.

4.3.2. RNAi and antisense oligonucleotides

The primary challenge in RNAi drug design lies in the target accessibility problem, where complex mRNA secondary and tertiary structures block hybridization sites. Traditional computational tools employed machine learning for efficacy prediction but

functioned primarily as screeners rather than generative designers[162,163].

A transformative approach has emerged through structure-based *de novo* design that fundamentally rethinks ASO−target interaction. The concept of "3D-ASOs" moves beyond Watson-Crick complementarity to recognize structured RNA targets through specific tertiary interactions including Hoogsteen pairs and base triples[156]. Molecular dynamics simulations provide mechanistic insights into these complex interactions, validating the importance of ASO length, triple base formation, and dynamic accessibility of structured regions[164]. This structure-aware paradigm enables AI models to design ASOs targeting previously undruggable structured RNAs, greatly expanding the therapeutic landscape for RNA interference.

4.3.3. Aptamers

The discovery of nucleic acid aptamers has traditionally relied on the laborious SELEX process, which explores only a minute fraction of theoretical sequence space. Generative AI now enables complete *in silico de novo* design workflows that bypass experimental selection entirely.

Structure-to-sequence approaches, exemplified by RhoDesign[123], take target 3D RNA structures as input and generate compatible nucleic acid sequences. Using geometric deep learning on structural scaffolds, these models generate novel aptamer sequences predicted to fold into conformations capable of target binding. Reinforcement learning strategies offer an alternative through fragment-based assembly. AiDTA[157] conceptualizes aptamer design as an iterative game where AI agents learn policies for selecting and assembling nucleic acid fragments, with rewards based on predicted binding affinity and structural similarity. Variational autoencoder-based approaches like DAPTEV[165] explore the latent space of aptamer sequences through evolutionary algorithms. These methods have successfully designed nanomolar-affinity aptamers from scratch, demonstrating cost-effective alternatives to traditional screening.

4.3.4. CRISPR-Cas systems

AI-driven design of CRISPR-Cas systems operates at two distinct levels: guide RNA optimization and complete system generation. For guide RNA design, deep learning models including DeepGuide[166] and SkipGuide[167] employ convolutional neural networks (CNN) and auto-encoders (AE) to learn complex mappings from sequence, genomic context, and epigenetic features to guide activity. These models enable prediction and ranking of optimal guide sequences for given targets, significantly improving genome editing reliability.

More profoundly, genomic foundation models enable *de novo* design of entire CRISPR systems. The Evo model[93], trained on whole prokaryotic genomes with 7-billion parameters, has learned co-evolutionary linkages between protein-coding and non-coding elements. When prompted, Evo generates novel kilobase-long DNA sequences coherently encoding both diverse Cas effector proteins and associated non-coding RNAs. This represents the ultimate form of *de novo* design, creating entire functional multi-molecular machines from first principles rather than optimizing individual components.

4.3.5. Current achievements and challenges in nucleic acid therapeutics

Nevertheless, critical challenges limit the translation of computational designs to clinical applications. The prediction of RNA dynamics and conformational ensembles remains less developed than for proteins, limiting the design of molecules requiring specific dynamic behaviors. Chemical modifications essential for nucleic acid drug stability and delivery are poorly represented in current models, which primarily focus on natural nucleotides. The cellular context profoundly influences nucleic acid function through interactions with RNA-binding proteins and processing machinery, yet these factors are largely absent from current design frameworks.

The fundamental challenge lies in bridging sequence−structure−function relationships in the complex cellular environment. While models can predict secondary structures with reasonable accuracy, tertiary structure prediction and the influence of cellular factors on RNA folding remain significant gaps. The design of chemically modified nucleic acids, essential for therapeutic applications, requires models that can handle non-natural nucleotides and predict their effects on stability, immunogenicity, and cellular uptake. Furthermore, the tissue-specific delivery of nucleic acid therapeutics remains a major bottleneck inadequately addressed by current AI approaches. Overcoming these challenges will require development of models that integrate chemical modifications, cellular context, and delivery considerations into the design process, alongside improved experimental platforms for rapid validation of computationally designed candidates.

5. AI-driven biologic drugs engineering

The development of a biologic drug faces multiple hurdles. Whether it is a natural molecule obtained through screening or an initial molecule discovered from scratch, few possess all the necessary characteristics for successful treatment directly. Lead candidates often require significant engineering to enhance its biological activity, improve stability, reduce immunogenicity, and optimize its pharmacokinetic profile[168]. This process of refinement is known as optimization. Traditionally, optimization has relied on labor-intensive methods like site-directed mutagenesis and random library screening[169]. However, the combinatorial explosion of possible sequence variants makes exhaustive experimental exploration infeasible. The emergence of AI, particularly advanced deep learning models[170], has revolutionized this landscape, enabling a paradigm shift from random exploration to intelligent, goal-oriented optimization. This chapter will first introduce the mainstream AI-driven methodologies powering this transformation and then delve into their specific applications in optimizing key classes of macromolecules, including antibodies, enzymes, and therapeutic peptides.

5.1. Methodological cornerstones: From zero-shot prediction to closed-loop learning

The power of AI in biological optimization stems from two complementary methodological paradigms that operate at different scales and with different data requirements. Understanding these foundational approaches is essential for appreciating how AI tackles diverse engineering challenges across molecular recognition, developability, and functional design.

5.1.1. Zero-shot optimization via pre-trained foundation models

Zero-shot optimization represents a paradigm shift in how we approach molecular engineering when experimental data is scarce or non-existent. The core insight is that the grammar of biological sequences—learned from analyzing millions of natural proteins

and nucleic acids—encodes fundamental biophysical constraints that are transferable to novel engineering contexts.

Protein language models (PLMs) like ESM-2[78], ProGen2[80], and ProtT5[171] trained on the entire UniProt database through self-supervised learning, have internalized evolutionary patterns that strongly correlate with stability and function. When tasked with optimizing a novel protein, these models can evaluate thousands of potential mutations by computing a pLL score—essentially asking "how natural does this sequence look?" A higher pLL suggests the mutation respects evolutionary constraints, providing a powerful filter before any experimental validation (Fig. 3C). While sequence-based PLMs are powerful, protein function is ultimately determined by its 3D structure. To bridge this gap, structure-based models like ESM-IF[114] and ProteinMPNN[116] were developed (Fig. 3D). These models take a backbone structure as input and predict a compatible amino acid sequence. By mutating specific residues and evaluating the model's confidence in the new sequence, they can predict the impact of mutations on structural integrity and stability.

The recent evolution toward multimodal foundation models has further expanded zero-shot capabilities. ESM3[84], which integrates sequence, structure, and functional annotations during training, can now perform structure-conditioned sequence optimization. Given a desired binding pocket geometry, it can propose sequences likely to adopt that configuration without requiring any experimental data for that specific fold. Similarly, genomic foundation models like Evo trained on 300 billion nucleotides with 131 kb context windows, have demonstrated zero-shot prediction of regulatory element function and RNA structure—capabilities that directly translate to optimizing mRNA therapeutics and gene editing tools.

The power and limitation of zero-shot approaches lies in their generality. While they excel at optimizing universal properties like stability that are well-represented in natural sequences, they struggle with highly specific objectives like binding to a synthetic antigen that has no evolutionary precedent. This gap motivates the second methodological pillar: data-efficient learning from experimental feedback.

5.1.2. Data-efficient optimization via model-in-the-loop directed evolution

When specificity matters—such as engineering an antibody against a novel viral variant or designing a peptide for a synthetic target—zero-shot methods must be augmented with experimental data. The challenge is that obtaining such data is expensive and time-consuming. Model-in-the-loop[172–174] approaches address this by creating a tight coupling between computational prediction and experimental validation, maximizing the information gained from each experiment.

Active learning exemplifies this paradigm (Fig. 3E). Starting with a small seed dataset (often fewer than 100 characterized variants), a surrogate model[172,175]—typically a Gaussian process or ensemble of neural networks—learns a local approximation of the fitness landscape. The key innovation is the acquisition function, which balances exploitation with exploration[176]. This transforms random screening into intelligent search; studies have shown 10–100 fold improvements in the efficiency of finding optimal variants compared to traditional directed evolution[175].

Advanced frameworks now integrate multiple machine learning techniques in sophisticated pipelines. The Low-N protein engineering approach combines variational auto-encoders (VAE) or other classic algorithm to learn a continuous latent representation of sequence space with Bayesian optimization to navigate it[177,178]. By operating in this learned latent space rather than raw sequence space, the optimization becomes smoother and more efficient—mutations that are close in latent space tend to have similar functions, enabling reliable gradient-based optimization even with sparse data.

5.2. Engineering molecular recognition: affinity and specificity enhancement

The ability to recognize and bind specific molecular targets with high affinity and exquisite selectivity forms the mechanistic basis for most biologic drugs. Whether it's an antibody neutralizing a virus, a peptide blocking a receptor, or a CRISPR system finding its genomic target, the engineering challenge is fundamentally about optimizing molecular recognition. AI has revolutionized this domain by learning the complex rules governing protein–protein, protein–nucleic acid, and protein–small molecule interactions.

5.2.1. Antibody affinity maturation through structure-informed learning

Therapeutic antibodies exemplify the molecular recognition challenge, where six CDRs must be optimized to achieve picomolar binding while maintaining stability[179,180]. Traditional affinity maturation through phage display is labor-intensive and explores only a minute fraction of sequence space. AI transforms this process through intelligent sequence design guided by structural principles.

A breakthrough approach by Shanker et al.[181] demonstrates the power of structure-informed language models. Using ESM-IF, which conditions its predictions on 3D structure, they evaluated antibody-antigen complexes to identify mutations that strengthen the binding interface. The model implicitly learns biophysical principles like optimal packing and electrostatic complementarity. Applied to COVID-19 neutralizing antibodies, this zero-shot method identified variants with 25-fold improved neutralization against viral escape mutants after testing fewer than 30 computationally selected candidates—a success rate that would require screening thousands of variants using traditional methods.

For more complex optimization involving multiple simultaneous mutations, Bayesian optimization frameworks have proven invaluable. Zeng et al.[182] developed a system that uses Gaussian processes to model the fitness landscape while a VAE proposes novel sequences in a learned latent space. This enables navigation of epistatic interactions—where the effect of one mutation depends on others—that confound single-mutation screens. The approach successfully evolved antibodies with 100-fold affinity improvements while maintaining thermostability, demonstrating that AI can balance the competing demands of binding and developability. These iterative AI-guided loops are critically enabled by high-throughput experimental platforms, such as the mammalian cell surface display and FACS screening systems described by Yang et al.[176], which provide the high-quality data needed to rapidly train and refine the predictive models in each cycle.

5.2.2. Peptide–target recognition optimization

Peptides present unique recognition engineering challenges due to their conformational flexibility and smaller binding interfaces compared to antibodies[183]. Recent AI approaches have moved beyond simple sequence optimization to consider the dynamic nature of peptide–target interactions.

The EvoBind[184] ingeniously repurposes AlphaFold2 as a fitness function for directed evolution. Given a target protein structure, it iteratively mutates a peptide sequence and selects variants that AlphaFold2 predicts will bind with high confidence, using metrics like pLDDT and pAE as proxies for binding likelihood. This structure-based approach has generated peptide binders for previously intractable targets, including intrinsically disordered proteins that lack defined binding pockets.

Meanwhile, generative language models are driving a paradigm shift from optimizing existing sequences to *de novo* design, which does not even require a 3D structure of the target. A representative example is PepMLM[185], which fine-tunes the ESM-2 protein language model to design novel linear peptide binders using only the target's amino acid sequence. By employing a masking strategy, PepMLM generates peptide sequences conditioned on the target sequence and has successfully designed specific binders for diverse disease targets, including those in cancer, Huntington's disease, and viral infections, showcasing the power of *de novo* design without structural input. Another breakthrough is AMP-Diffusion[186], a generative diffusion model that operates in the latent space of protein language model embeddings to create entirely new anti-microbial peptides (AMPs). By learning and denoising within this latent space, the model generates a vast number of candidates with broad-spectrum antibacterial activity and low toxicity, whose efficacy, comparable to standard antibiotics, has been validated *in vivo*.

These advanced optimization frameworks are critically dependent on high-quality predictive models for various peptide properties. Foundational models like PepLand[187] address a key gap by providing rich representations specifically for peptides, including those with nAAs. Unlike general protein language models, PepLand is built on a heterogeneous graph neural network that captures atomic and fragment-level features, showing superior performance in predicting properties like solubility and permeability. In parallel, specialized models are being developed for specific applications. For AMPs, LLAMP[188] fine-tunes the ESM-2 language model to be species-aware, predicting the minimum inhibitory concentration (MIC) of a peptide against a specific bacterium by incorporating genomic features of the target species. This allows for the targeted optimization of AMPs against high-priority pathogens, moving beyond generic activity prediction to achieve species-specific potency. Together, these multi-faceted AI strategies are transforming peptide engineering from a trial-and-error process into a rational, goal-driven design cycle.

5.2.3. *Precision enhancement in gene editing systems*

The specificity of CRISPR-Cas systems—distinguishing between intended genomic targets and near-identical off-target sites—represents a critical molecular recognition challenge for therapeutic applications[189]. AI has enabled systematic optimization of both guide RNAs and Cas proteins to enhance on-target activity while minimizing off-target editing[190,191].

For guide RNA design, deep learning models like DeepGuide[166] use CNN to learn the complex relationship between guide sequence, genomic context and editing efficiency. Trained on genome-wide CRISPR screens, these models can predict and rank thousands of potential guides, identifying those with optimal on-target activity and minimal off-target potential. At the protein level, AI-guided evolution of Cas enzymes has yielded variants with dramatically improved specificity. He et al.[192] leveraged pre-trained PLMs to score thousands of *in silico* variants of a uracil-*N*-glycosylase (UNG), an enzyme not typically used in base editing. By testing fewer than 50 top-ranking variants, they successfully identified an enhanced UNG that, when fused to nCas9, created TSBE3, a novel and efficient base editor for T > G/C substitutions. This remarkable success, achieved with minimal wet-lab work, demonstrates the profound capacity of AI to navigate sequence space and generate novel, highly active enzymes for bespoke gene editing applications.

5.2.4. *De-immunization and humanness engineering*

For most therapeutic proteins, minimizing immunogenicity is a paramount developability objective to prevent the generation of anti-drug antibodies (ADAs). AI is transforming the traditionally empirical process of de-immunization into a predictive, goal-oriented engineering task through two primary strategies: T-cell epitope removal and structure-guided humanization.

The first strategy focuses on identifying and eliminating T-cell epitopes—short peptide sequences presented by MHC molecules that can trigger an immune response. AI excels at predicting peptide–MHC binding to identify these immunogenic hotspots. Advanced architectures, such as the capsule neural network in CapsNet-MHC[193], enhance prediction accuracy by modeling the hierarchical features of the entire peptide–MHC complex, leading to more reliable epitope identification across genetically diverse populations. This allows engineers to then propose silent mutations that remove immunogenicity while preserving the protein's function.

For non-human antibodies, AI enables a paradigm shift from conventional CDR grafting to more precise, structure-guided paratope grafting. Traditional methods often fail to preserve affinity and can retain unnecessary foreign sequences. To overcome this, AI models like XtalFold[194] predict the high-resolution 3D structure of the antigen–antibody complex. This structural insight facilitates the superior technique of transplanting only the essential antigen-contacting residues (the paratope)—which may reside in both CDR and framework regions—onto an optimal human framework. As demonstrated by Wang et al.[194], this approach can rescue a failed humanization attempt by identifying critical non-CDR residues, achieving a level of precision that minimizes immunogenicity risk while crucially preserving binding affinity.

5.3. *Optimizing biophysical properties for developability*

A biologic drug must not only bind its target but also survive the journey from manufacturing to patient administration. This requires optimization of multiple biophysical properties—stability, solubility, viscosity, and aggregation resistance—that determine whether a molecule can be produced at scale, formulated at high concentrations, and remain active during storage. AI has emerged as a powerful tool for engineering these developability properties, often revealing non-obvious solutions that enhance multiple properties simultaneously.

5.3.1. *Engineering protein stability*

Protein aggregation represents a critical failure mode for biologics, potentially causing immunogenicity and loss of efficacy. AI transforms traditional stability engineering by learning the sequence determinants of stability from large-scale datasets. A multi-tiered AI strategy has emerged, progressing from rapid zero-shot screening to highly accurate, evolution-informed predictions.

A powerful starting point is zero-shot prediction, where pre-trained models are applied directly without task-specific training. As systematically evaluated by Reeves et al.[195], many protein

sequence likelihood models can effectively rank mutants by stability, with structure-based models like ProteinMPNN providing a robust baseline. While these zero-shot methods offer broad applicability, predictive accuracy can be significantly enhanced through fine-tuning on experimental data. Here, the choice of model architecture is paramount; models incorporating evolutionary context[196], such as the MSA Transformer, achieve state-of-the-art performance for single-point mutations, underscoring that co-evolutionary signals are vital for precise predictions.

The challenge escalates with multiple mutations due to epistasis. Counterintuitively, simple additive models are often surprisingly effective for predicting the stability of double mutants in some cases. However, research also reveals a crucial exception[197]: models explicitly designed to capture non-additive effects provide a distinct advantage in the most critical task for engineers—predicting and identifying synergistic stabilizing mutations. This illustrates a sophisticated approach to stability engineering, leveraging different AI tools tailored for specific levels of mutational complexity.

5.3.2. Optimizing nucleic acid stability and expression

For nucleic acid therapeutics, engineering focuses on optimizing stability against nucleases and, for mRNA, maximizing translational efficiency. AI provides powerful tools to navigate this vast and complex design space.

One transformative strategy for mRNA, exemplified by LinearDesign[198], reframes sequence design as a computational linguistics problem. It uses an efficient algorithm to parse the astronomical number of synonymous coding sequences, identifying the single sequence with the lowest minimum free energy while co-optimizing for codon usage. This dual optimization yielded dramatic results: mRNA vaccines designed by this method showed up to 128-fold higher antibody titers *in vivo* compared to conventional designs. Complementary tools like mRNAdesigner[199] modularize the process, using AI to separately optimize the CDS, 5′ UTR, and 3′ UTR for factors like GC content and ribosome loading, and then integrating them into a final high-expression construct.

The fundamental reason these codon optimization strategies are so effective is that codon choice is far more than a simple redundancy. Work by Outeiral et al.[200] demonstrates that language models trained on codon sequences capture significantly more biological information than those trained on amino acid sequences alone. Their codon adaptation language model outperformed similarly sized protein-only models in predicting properties like protein stability and abundance, proving that synonymous codon usage is a rich, underutilized signal that reflects deep evolutionary and biophysical constraints. This provides a strong theoretical underpinning for why AI-driven codon optimization is a cornerstone of modern mRNA engineering.

5.3.3. Enhancing intrinsic delivery properties and bioavailability

For many biologics, crossing biological membranes is a fundamental barrier to reaching intracellular targets. AI is enabling the systematic engineering of cell-penetrating properties and the optimization of delivery formulations.

At the molecular level, frameworks like PepTune[201] use multi-objective discrete diffusion models to design peptides that balance target binding with membrane permeability. By operating on SMILES representations, these models can explore non-natural amino acids and cyclization, generating candidates that maintain nanomolar target affinity while possessing cell-penetrating properties. For oligonucleotide drugs, tools like ASOptimizer[202] use graph-transformers to guide the strategic placement of chemical modifications (*e.g.*, 2′-*O*-methyl) that enhance cellular uptake and nuclease resistance, computationally screening thousands of patterns to find an optimal balance.

At the formulation level, AI is accelerating the optimization of delivery vehicles like LNPs. Platforms such as AGILE[203] use machine learning to predict transfection efficiency from lipid structure and formulation parameters. This allows for vast virtual screening, accelerating the development of next-generation delivery systems with enhanced tissue targeting and reduced toxicity.

5.4. Designing novel functions: Catalysis, regulation, and complex architectures

Beyond optimizing existing functions, AI enables the engineering of entirely new biological capabilities—from novel catalytic activities to programmable regulatory switches. This represents the frontier of biologic drug engineering, where AI doesn't just refine but fundamentally expands what biological molecules can do therapeutically.

5.4.1. Engineering to discover new catalytic functions

The design of therapeutic enzymes with novel or enhanced catalytic activities has long been a holy grail of protein engineering. AI has transformed this challenge by learning the sequence-structure-function relationships that govern catalysis.

A powerful AI-driven strategy to achieve this has moved beyond modifying known enzymes to the *de novo* discovery of entirely new catalysts from nature's vast sequence and structural databases. The field of gene editing offers a prime example of this approach, where programmable nucleases and deaminases act as therapeutic catalysts. However, pioneering systems like CRISPR-Cas9 face practical limitations, such as their large size, which complicates delivery using vehicles like AAV vectors. AI is accelerating the discovery and optimization of novel, hyper-compact nuclease systems to overcome this limitation. Using structure-based protein clustering with AlphaFold2 predictions, Huang et al.[204] mined entire deaminase families to discover previously unknown functional relationships. This led to the identification of a suite of novel single-strand-specific DNA deaminases, which are not only smaller but also show increased activity and minimal off-targets. Critically, some of these adds enabled robust base editing in soybean, a crop previously recalcitrant to such technologies, and their compact size makes them ideal for AAV packaging. In a parallel effort, Yoon et al.[205] employed a structure-first search pipeline combining the speed of Foldseek with the sensitivity of DALI to uncover Cas13an, a deeply divergent, ancestral, and compact RNA nuclease one-third the size of its relatives, highlighting the power of structure-guided mining to explore remote protein families.

5.4.2. Creating programmable molecular switches

The ability to control protein function in response to external signals is crucial for next-generation therapeutics. AI has enabled the systematic design of allosteric switches and logic gates within proteins.

The ProDomino[206] model addresses the fundamental challenge of identifying sites within proteins that tolerate domain insertion without destroying function. Trained on a large dataset of natural domain insertion events, it predicts permissive sites with high

accuracy. This enabled the creation of light- and chemically-inducible versions of CRISPR-Cas9 and Cas12a, where effector activity is controlled by inserted photoreceptor or ligand-binding domains. The INSRTR platform[207] extends this concept to create molecular logic gates. Using gradient boosting models trained on experimental insertion data, it guides the insertion of peptide actuators into target proteins to create AND, OR, and NOT gates within single polypeptides. This enables sophisticated therapeutic responses—for example, a CAR-T cell that activates only in the presence of multiple tumor antigens, dramatically improving specificity.

5.4.3. *Optimizing multi-component therapeutic systems*

Many advanced biologics comprise multiple functional components that must work in concert—such as therapeutic fusion proteins[95], bispecific antibodies, and enzyme cascades. AI excels at optimizing these complex architectures where function emerges from the interaction between components.

The iMARS[208] framework revolutionizes the design of enzyme cascades by introducing a space-efficiency code that quantifies optimal spatial arrangements. Using AlphaFold2 to model thousands of fusion constructs with different linkers, it computes scores based on inter-active site distances and orientations. This systematic approach yielded fusion enzymes with 45-fold improved activity for metabolite production, demonstrating how AI can optimize the emergent properties of multi-enzyme systems. For ADC, which combine antibodies, linkers, and cytotoxic payloads, AI models like ADCNet[209] integrate information from all components to predict overall therapeutic activity, guiding the rational design of next-generation conjugates with improved therapeutic windows.

5.4.4. *Engineering immune responses: rational vaccine design*

In stark contrast to de-immunization, the goal of vaccine design is to engineer antigens for enhanced immunogenicity—specifically, to elicit a potent and broadly neutralizing antibody (bnAb) response. AI serves as a critical tool for rational immunogen design, strategically guiding the immune system toward desired targets. A primary application is epitope focusing, where machine learning models analyze viral sequence and structural data to identify conserved, vulnerable epitopes targeted by known bnAbs[210]. Generative models like ProteinMPNN can then be employed to redesign the antigen, for instance, by removing immunodominant but non-neutralizing decoy epitopes to redirect the immune response. Furthermore, AI-powered structural models can rapidly screen for mutations that lock an antigen in a specific conformation (*e.g.*, the pre-fusion state of viral glycoproteins) that is known to be the primary target of potent antibodies, a strategy famously used in the development of COVID-19 vaccines[211]. By combining epitope focusing and conformational stabilization, AI enables the creation of next-generation immunogens that are intelligently optimized to elicit a superior protective immune response.

5.5. *Overarching challenges and future directions in AI-driven optimization*

While AI excels at optimizing single parameters like binding affinity or thermal stability, the central challenge in modern biologics engineering lies in multi-parameter optimization (MPO). A successful therapeutic requires a delicate balance of numerous, often conflicting, properties including efficacy, stability, solubility, low immunogenicity, and favorable pharmacokinetics. The primary bottleneck to effective MPO is the scarcity of predictive models for complex *in vivo* properties, which in turn creates a cold start problem for data-efficient, model-in-the-loop learning campaigns that require initial experimental data. While zero-shot foundation models provide a powerful starting point for general properties well-represented in evolutionary data, their reliability diminishes for highly specific or novel functions. This underscores the core challenge: intelligently navigating a high-dimensional and data-sparse landscape to find rare candidates that satisfy multiple clinical and manufacturing constraints simultaneously.

Perhaps the most fundamental limitation, however, is the persistent gap between *in vitro* optimization and *in vivo* efficacy. Molecules that are computationally perfected and display ideal characteristics in simplified lab assays often fail within the complex, dynamic system of a patient. Bridging this gap requires a strategic shift beyond simply scaling current methods. It demands the generation of richer, multi-modal datasets that capture *in vivo* complexity, and the development of a new generation of AI models. These future models must move beyond predicting static molecular properties to account for pharmacokinetics, systems-level interactions, and dynamic biological processes. Successfully closing this *in vitro*-to-*in vivo* gap is the most critical step toward realizing the vision of closed-loop, autonomous systems that can rapidly and reliably translate computational designs into clinically successful therapeutics.

6. AI-accelerated the development of macromolecules delivery systems

Efficient and targeted *in vivo* delivery of macromolecules presents a critical bottleneck for their therapeutic efficacy. AI is revolutionizing the design, optimization, and evaluation methods of biological molecular delivery systems, marking a significant shift from traditional trial-and-error approaches to a rational, intelligent design paradigm. By adeptly learning complex data patterns and deciphering intricate structure−function relationships, AI assumes a central role in the intelligent design of these systems, thereby substantially accelerating the discovery and optimization of novel delivery technologies.

This transformative shift is illustrated in Fig. 4. As shown in Fig. 4A, discriminative and generative AI models play complementary roles in advancing delivery systems such as LNPs, AAV, and ADC. Discriminative models function as predictive evaluators, forecasting key properties including biodistribution, stability, and immunogenicity. This capability enables accurate *in silico* assessment, guiding high-throughput screening and prioritizing promising candidates, thereby substantially reducing experimental costs. In practice, these models have been successfully applied to predict the transfection efficiency of ionizable lipids in LNPs, assess the therapeutic efficacy of ADC, and predict the tissue specificity of AAV capsids. Meanwhile, generative models facilitate *de novo* molecular design, enabling the creation of optimized ionizable lipids for LNPs, the design of linker chemistries for ADC, and the engineering of viral capsid variants for AAV. Importantly, generative models can be integrated with discriminative predictors to establish closed-loop optimization cycles, simultaneously expanding the design space and enhancing predictive reliability.

The advantages of this AI-driven approach become clear when contrasted with traditional methods, as detailed in Fig. 4B.

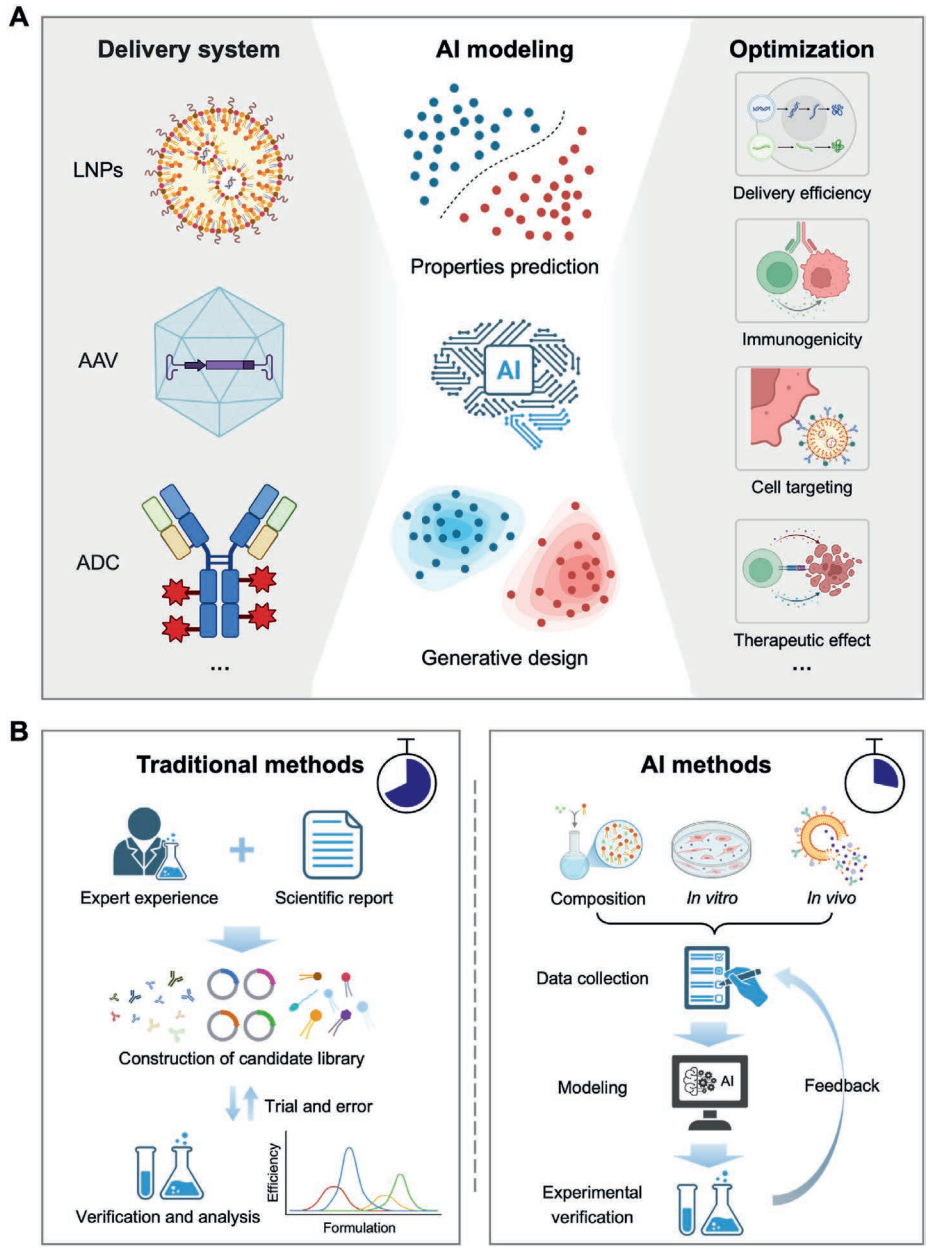

Figure 4 AI-driven optimization of delivery systems. (A) The AI optimization paradigm for drug delivery systems. Application of artificial intelligence in improving delivery systems such as LNPs, AAV, and ADC. Discriminative models predict key properties from existing data, while generative models design novel systems. This dual approach enables the optimization of critical performance metrics, including delivery efficiency, reduced immunogenicity, and specific cell targeting. (B) Comparison of AI-driven and traditional methods for delivery systems design. A contrast between the iterative, data-centric AI workflow and the time-consuming, trial-and-error approach of traditional methods.

Traditional workflows rely heavily on expert intuition and labor-intensive, iterative experimentation, which is time-consuming and limits the exploration to a narrow design space. In sharp contrast, the AI-driven workflow systematically integrates data collection, predictive modeling, and experimental verification in a closed-loop manner. This synergy of discriminative and generative modeling reduces experimental burden, accelerates optimization, and allows for the exploration of a much broader design space.

Collectively, the integration of AI establishes a powerful paradigm for the rational engineering of delivery systems, offering a transformative pathway toward efficient, safe, and clinically viable macromolecule delivery. In the following subsections, we will discuss specific applications of these AI approaches and their potential to address key challenges in delivery system design.

6.1. AI-driven LNPs optimization

The remarkable success of lipid nanoparticles as delivery vehicles for COVID-19 mRNA vaccines (*e.g.*, mRNA-1273 and BNT162b2) in eliciting immune responses against SARS-CoV-2 infection has profoundly stimulated interest in LNPs delivery systems and their broad application potential[50,51]. However, a

highly complex and non-linear relationship exists between LNPs formulation (including the composition and ratio of lipids) and their corresponding physicochemical properties (such as particle size, zeta potential, and polydispersity index). Furthermore, these properties are intricately linked to subsequent biological functions (*e.g.*, encapsulation efficiency, cellular uptake, *in vivo* biodistribution, and immunogenicity). Traditional experimental optimization methods mainly rely on combinatorial library construction and high-throughput screening; though helpful, they are inherently inefficient and costly. In recent years, researchers have actively explored leveraging machine learning methods to predict various LNPs properties, particularly transfection efficiency, which is the functional basis for the further optimization of LNPs. By learning intricate patterns from vast datasets of LNPs formulations and experimental settings, these models aim to construct prediction models of LNPs formulation−function relationships, thereby accelerating the exploration of the expansive LNPs parameter space and enabling the design and optimization of LNPs formulations with target properties.

Wang et al. utilized ECFPs to encode the molecular structures of ionizable lipids. They developed the LightGBM gradient boosting framework, which demonstrated excellent performance in predicting LNPs related properties and offered a certain degree of interpretability, facilitating insights into the relationships between molecular structures and biological function[212,213]. Ding et al.[214] systematically compared traditional rule-based molecular representations (*e.g.*, RDKit descriptors) with deep molecular representations derived from a pretrained graph neural network model (*e.g.*, Grover) for transfection efficiency classification tasks. They evaluated various downstream classifiers, finding that the combination of RDKit descriptors with the multilayer perceptron achieved the best performance on their test set. Moayedpour et al.[215] explored the potential of large language models, specifically the MolBART model, in modeling lipid molecules. Their study showed that MolBART achieved high prediction accuracy and strong generalization ability, even when finetuned with limited labeled samples.

The AGILE platform[203] proposed a feature fusion framework that integrates a pretrained molecular graph encoder with handcrafted molecular descriptors to comprehensively represent lipid structures, which was successfully applied to both transfection efficiency prediction and virtual screening of ionizable lipids with high transfection potential. Witten et al.[216] curated a substantial dataset of over 9000 LNPs activity measurements and developed a DMPNN model based on the dataset; it showed promising performance in both transfection efficiency prediction and ionizable lipid virtual screening. TransLNP, built upon the three-dimensional molecular pretrained model Uni-Mol, innovatively integrates coarse-grained atomic sequence information and fine-grained atomic spatial relationship data of ionizable lipids, incorporating both label and feature distribution smoothing techniques to alleviate the problem of limited experimental data and imbalanced sample distribution[217]. By deeply excavating the complex mapping between lipid structure and transfection efficiency within Transformer architecture, the model consistently demonstrates superior performance across various task settings.

To overcome the limitations of single-molecule-based algorithms, which are insufficient for capturing the composite nature of LNPs, Chan et al.[218] developed COMET (composite material Transformer), a flexible framework that jointly encodes molecular structures, molar ratios, and synthesis parameters to holistically represent LNPs formulations. Trained on the curated LANCE dataset comprising over 3000 formulations, COMET achieved accurate prediction of excluded samples and demonstrated strong generalizability by forecasting LNPs performance in novel cell lines as well as stability after lyophilization, even with limited training data. Moreover, by conducting an *in silico* screening of 50 million virtual LNPs, COMET identified top candidates with experimentally validated *in vitro* and *in vivo* protein expression, thereby providing a scalable computational strategy to accelerate nucleic acid delivery system development.

Generative artificial intelligence has also been used in the molecular design of LNPs. Ou et al.[219] innovatively proposed a deep generative model for the design of easily synthesized ionizable lipids. This method first screened lipid head and tail building blocks that met the molecular weight, solubility, and ionizability characteristics from the ZINC20 database. Based on the improved Synthesis-DAGs framework, a directed acyclic graph (DAG) was used to represent the synthesis path. The action sequence, including node addition, building block selection, and connection mode, was modeled by a recursive neural network, and the chemical reaction process was accurately simulated in combination with the Chemformer reaction predictor. Experimental results show that 92.6% of the molecules generated by this method are effective lipids, 83.4% have ionization characteristics, and show excellent uniqueness and novelty. Iterative optimization combined with the property prediction model can significantly improve the target properties of the generated ionizable lipids. This study expands the design space of ionizable lipids while maintaining the synthesizability of the molecules by combining data-driven building block screening, structured synthesis path generation, and accurate reaction prediction.

Meanwhile, Zhao et al.[220] proposed a deep generative model based on MCTS for the lipids design, which utilizes a synthetically accessible lipid building block dataset and two specialized predictors to guide the search through the chemical space. The MCTS generative model is guided by a policy network, where the latter is trained by maximizing its similarity to the visit count distribution. Experimental results indicate that some of the lipids generated by this method possess corresponding synthetic pathways, and these pathways have been validated. This research expands the design space for ionizable lipids while maintaining molecular synthesizability by integrating data-driven building block screening, structured synthesis path generation through MCTS exploration, and precise reaction prediction.

6.2. *AI-guided AAV sequence design and optimization*

AAV has emerged as a leading vector for gene therapy due to its broad host range, low pathogenicity, and ability to facilitate long-term gene expression. However, AAV vectors face limitations, including immunogenicity and insufficient targeting specificity. To improve the efficacy of gene therapy, biologists commonly employ directed evolution, which involves introducing random mutations in specific amino acid regions of the AAV capsid protein, and then evaluating their impact on AAV delivery characteristics to iteratively enhance. However, this experimental approach for optimizing AAV vector is often time-consuming, resource-intensive, and difficult to replicate. The advancement of AI, particularly machine learning, offers significant potential to accelerate capsid optimization while reducing development time and manufacturing costs.

Ogden et al.[221] adopted a machine-guided design approach to capsid protein engineering, performing saturation mutagenesis on

all 735 amino acid positions within the AAV2 capsid, including all possible codon substitutions, insertions, and deletions at each site. This generated a library of approximately 200,000 viral variants, from which mutants capable of both maintaining AAV2 viability and improving its homing potential (*i.e.*, tropism) to specific organs in mice were identified. Building on this, the researchers proposed an additive model constructed from single-mutation data, which can approximate the fitness of multi-mutant variants, thereby enhancing design efficiency. Specifically, the machine learning framework sampled mutations at each position within the capsid protein's 561 to 588 region (which includes both surface-exposed and buried residues) proportionally to their measured effect on liver delivery, subsequently designing multi-mutant variants. These designed variants were then tested alongside randomly generated mutants in liver biodistribution experiments in mice. The results confirmed that, compared to traditional random mutagenesis, machine learning-guided design significantly increased the proportion of functional liver-targeting mutants, with its advantage becoming particularly pronounced when introducing four or more mutations. This research powerfully demonstrates the immense potential of machine learning in systematically optimizing AAV capsids, imbuing them with enhanced properties for emerging gene therapies.

In addition to optimizing specific organ tropism, AI also demonstrates its ability to improve AAV vector packaging efficiency. Bryant et al.[222] leveraged deep learning to design highly diversified AAV2 capsid protein variants capable of efficiently packaging DNA payloads. Focusing on a specific 28-amino acid segment of the AAV2 wild-type capsid protein, the researchers generated and assayed a vast number of variants *via* high-throughput experiments. This functional activity data was then directly used to train various predictive models, including simple logistic regression (LR), CNN, and recurrent neural network (RNN). These models demonstrated precise predictive capability for capsid DNA packaging viability, even when trained on limited datasets, thereby effectively guiding the intelligent exploration of vast sequence spaces. Through model prediction and iterative design, the method generated a total of 201,426 capsid variants, of which 110,689 were confirmed to be biologically viable. Notably, 57,348 of these viable variants exhibited sequence diversity surpassing the average diversity of natural AAV serotype sequences, with these highly diversified active variants carrying an average of 12 to 29 mutations within the 28-amino acid segment. This achievement not only demonstrates AI's capacity to efficiently unlock and generate functional protein sequences within deep sequence spaces without relying on complex biophysical models, but also lays a solid foundation for the rational design and rapid optimization of future viral vectors and protein therapeutics.

Similarly, Zhu et al.[223] constructed a standard NNK 7-mer peptide insertion experimental library and analyzed the packaging fitness of each variant (expressed as logarithmic enrichment score) by high-throughput sequencing technology. Based on this dataset, they trained a neural network to predict the relationship between peptide sequence and packaging efficiency. In order to optimize the library design, the researchers set sequence diversity constraints (measured by entropy) for the inserted peptides and generated a series of Pareto optimal libraries by balancing the diversity of the library and the average predicted packaging efficiency. Experimental verification shows that machine learning optimized libraries show significant advantages over traditional NNK libraries: the packaging titer is increased by 5 times while maintaining more than 90% sequence diversity. More notably, in the screening of primary human brain tissue infection, the number of effective variants in the optimized library reached 10 times that of the NNK library, and highly efficient variants with glial cell targeting specificity were successfully identified.

Optimization of a single characteristic of AAV capsid cannot fully meet the needs of effective delivery; it is often necessary to optimize multiple characteristics simultaneously in practice. However, the vast AAV capsid sequence space and the inefficiency of traditional screening methods in multi-trait optimization have limited the discovery of clinically viable vectors. To address this, the Fit4Function platform proposes a generalizable machine learning approach for systematically engineering multi-trait AAV capsids, moving beyond conventional limitations[224]. Its core involves constructing a meticulously designed Fit4Function capsid library that uniformly samples the manufacturable sequence space. Leveraging this library, researchers generated high-quality, reproducible experimental data through extensive *in vitro* (*e.g.*, human cells) and *in vivo* (*e.g.*, mouse) screens. This data was then used to train accurate sequence-to-function ML models, enabling simultaneous optimization of multiple traits by integrating ML models for properties like liver tropism and manufacturability within a unified framework. Experimental results demonstrate that Fit4Function successfully designed a multi-trait optimized AAV capsid library, with 88% of variants meeting predefined evaluation criteria, confirming its AI-driven capability to efficiently generate AAV capsids possessing multiple desired characteristics. Crucially, despite models being trained solely on mouse *in vivo* and human *in vitro* data, they accurately predicted AAV capsid variant biodistribution in macaque (non-human primate) organs. This powerful cross-species generalizability represents a significant breakthrough, substantially accelerating AAV vector development by enabling early prediction of performance in complex primate models and humans from more accessible data, thereby drastically reducing costly and time-consuming large-scale experiments.

Generative artificial intelligence has also been used in the design of AAV capsid sequences to meet the requirements of multi-functional relationships. Guo et al.[225] proposed a machine-learning design method to map an AAV capsid sequence to multiple functions, thereby enabling direct *in silico* engineering of AAV capsids. To fuse multiple functions into a single capsid sequence, a heuristic algorithm coupled with contrastive learning and reinforcement learning, named function-guided evolution, was introduced to steer further evolution of the high-performing capsid sequences generated by a naïve language model toward functions. Despite the inherent constraint of relatively small datasets, with the study leveraging only 129 entries, the researchers successfully constructed a robust model capable of mapping AAV capsid sequences to multiple functions, specifically demonstrating improved viability coupled with enhanced central nervous system (CNS) tropism. The effectiveness of this *in silico* design was subsequently confirmed through *in vivo* experiments. These experiments revealed that two of the top eight engineered variants exhibited both enhanced viability and remarkable CNS tropism. This interpretable generative artificial intelligence design method signifies a pioneering effort, establishing a direct computational

pathway for engineering AAV capsids for highly effective delivery applications.

6.3. Relevant applications of AI in ADC

ADC represent a revolutionary therapeutic modality in the era of precision medicine, offering immense potential for cancer treatment by specifically targeting cancer cells and releasing highly potent drugs. However, the rational design and discovery of ADC face a core challenge: the highly complex relationship between their five key components: target antigen, antibody, linker, payload, drug−antibody ratio (DAR), and the ultimate drug activity is exceedingly difficult to predict accurately.

To address this complexity, Chen et al.[209] proposed ADCNet, a unified deep learning framework, which comprehensively integrates and profoundly analyzes information from all critical ADC components and their intricate interactions. Specifically, it employs advanced pre-trained models, such as ESM2 for efficient feature extraction from protein sequences (including both target antigen and antibody), and FG-BERT for feature extraction from small molecules (*i.e.*, the payload and linker). Features of different components are subsequently integrated through a feature fusion framework, ultimately enabling the precise prediction of ADC's biological activity. Through this data-driven and highly intelligent methodology, ADCNet is poised to significantly accelerate and optimize the rational design and discovery process for next-generation ADC, overcoming the limitations of traditional trial-and-error approaches and substantially enhancing ADC development efficiency and success rates.

The relevant property prediction framework is also used in the screening of payload molecules in ADC, Guo et al.[226] established a graph attention network (GAT)-driven dual scoring model, designed to simultaneously predict drug payload membrane permeability and cytotoxicity, which was utilized for constructing novel ADC with efficient bystander killing effect. This study first engaged in the rational design of exatecan derivatives (Eds), through which the researchers discovered that the cyclobutyl-modified Ed9 exhibited both excellent permeability (showing a 3-fold increase over DXd in PAMPA experiments) and potent anti-tumor activity (with a 2- to 10-fold reduction in IC_{50} values). This led to the successful construction of a homogeneous ADC, which demonstrated a 40% higher killing efficiency against HER2-negative cells than T-DXd in *in vitro* experiments, and achieved complete tumor regression in xenograft models. This work innovatively integrates AI prediction with chemical biology validation, thereby offering a new ADC design paradigm for overcoming tumor heterogeneity.

Xu et al.[227] proposed DumplingGNN, a novel hybrid GNN architecture designed for predicting the activity of ADC's payloads. DumplingGNN integrates MPNN, GAT and GraphSAGE layers to achieve multi-scale feature extraction and enhance molecular representations with crucial 3D structural information, thereby significantly boosting prediction performance. On a customized dataset of topoisomerase I inhibitor ADC payloads, the model demonstrated high accuracy in classification prediction. Furthermore, the model's attention mechanism provides valuable interpretability, enabling the identification of pharmacophore features directly correlated with activity, thus offering biological insights for rational drug design.

In the synthesis of ADC, the high variability of bioconjugation reaction outcomes, particularly the unpredictability of the drug−antibody ratio, constitutes a limiting factor that impacts biological activity. To address this bottleneck, Angiolini et al.[228] developed a machine learning framework built upon their laboratory's proprietary dataset of ADC-related experimental results, which integrates information pertaining to the experimental system, mAb characteristics, conjugation type, and the molecular structure of the linker-payload to predict the DAR value of ADC utilizing an XGBoost regression predictor. This work provides a crucial tool for the rational design of ADC, enabling researchers to predict preliminary experimental results for a given system, optimize components and synthesis conditions *in silico* before undertaking costly wet-lab experiments, thereby significantly accelerating the ADC development process.

Generative AI models, which enable the *de novo* design of new molecular entities, are increasingly applied in ADC development to expand the constrained design space of chemical linkers that are critical for stability and efficacy. Su et al.[229] introduced Linker-GPT, a Transformer-based deep learning framework, which leverages self-attention mechanisms to generate novel ADC linkers with high structural diversity and synthetic feasibility. The framework operates by integrating transfer learning from large-scale molecular datasets with reinforcement learning to iteratively refine molecular properties such as drug-likeness and synthetic accessibility, which provides a scalable computational platform for accelerating the discovery and optimization of novel ADC linkers, thereby overcoming the limitations of traditional, narrow design motifs.

7. Overarching challenges and strategic frontiers

While the integration of AI has substantially expanded the exploration of chemical and biological space, a critical discordance remains: the exponential growth in computational capabilities has not yet translated into a linear improvement in clinical success rates. The field is currently transitioning from a phase of feasibility demonstration which involves showing that molecules can be generated to clinical engineering where the focus shifts to efficacy and safety. This transition is constrained not by computational power, but by structural limitations in data quality, the biological complexity of systemic interactions, and the interpretability required for regulatory compliance. The following sections outline the strategic frontiers that must be addressed to bridge the gap between *in silico* predictions and therapeutic reality.

7.1. From data scarcity to an AI-native experimental ecosystem

The performance of AI models depends critically on data quality and relevance. The initial breakthroughs were built upon large, publicly available, but ultimately general-purpose datasets (*e.g.*, UniProt, PDB). Subsequently, AlphaFold's 214 million predicted structures significantly supplemented these existing datasets[230]. The current limitation is not data volume, but the scarcity of high-quality, task-specific labeled data, and multi-modal labeled data required for fine-grained therapeutic optimization.

As we saw in the sections on optimization, public databases are inherently biased towards stable, well-characterized proteins and critically lack standardized negative data—information on variants that fail—which is essential for teaching AI to identify design constraints. Additionally, the cost of labeling multiple developability properties such as stability and immunogenicity across identical molecules or delivery systems creates a significant data bottleneck.

The strategic imperative, therefore, is a shift from passively using archival data to actively and intelligently generating AI-

native datasets. This is being realized through the development of closed-loop, robotic platforms that automate the "design-build-test-learn" (DBTL) cycle. In these systems, active learning algorithms guide experiments to maximize information gain per assay, establishing a closed-loop system where experimental results continuously refine computational models. The future lies in building a robust ecosystem where these automated platforms generate standardized, multi-modal data—linking sequence not just to structure, but directly to functional outcomes, pharmacokinetic profiles, and delivery efficiencies—thereby providing the high-fidelity input needed for next-generation AI models.

7.2. *Bridging the molecular-to-system complexity discrepancy*

A critical challenge is the discrepancy between molecular-level predictions and systemic biological outcomes. Current models achieve high accuracy in predicting static properties, such as protein folding or binding affinity. However, a biologic drug's success is ultimately determined by its behavior within the vastly more complex biological system of a patient.

This "*in silico-in vivo*" disconnect persists across development stages. For instance, a *de novo* designed binder may possess a high confidence score yet exhibit poor expression or solubility. An antibody, optimized for picomolar affinity in a buffer solution, may exhibit poor pharmacokinetics or off-target toxicities *in vivo*. A lipid nanoparticle, proven effective for transfection in a cell culture dish, may be cleared by the immune system or fail to target the intended organ. Similarly, targeted delivery systems such as AAV capsids and ADC, which are optimized for tissue specificity and payload release in preclinical models, may exhibit altered biodistribution, reduced efficacy, or unexpected immune responses in patients. Resolving this complexity is crucial. It requires a new class of AI models that can learn to integrate multiple scales of biological organization. Future models must integrate conformational dynamics and cellular context—accounting for interactions with chaperones, proteases, and the immune system. This frontier will likely depend on the fusion of data-driven deep learning with physics-based simulations and quantitative systems pharmacology models, creating hybrid approaches that honor both biological data and fundamental physical laws.

7.3. *The imperative for controllability and interpretability*

As model complexity increases, issues of controllability and interpretability become critical. Generative models enable the exploration of extensive chemical spaces; however, this capability must be constrained to meet therapeutic requirements.

Multi-parameter optimization exemplifies the need for precise controllability. Methodologies must advance beyond *post-hoc* filtering of generated libraries. The future requires more sophisticated conditional generative models and reinforcement learning frameworks where an AI agent learns an optimal policy for making sequential modifications to a molecule to navigate the complex fitness landscape and arrive directly at Pareto-optimal solutions.

This need for control is intrinsically linked to interpretability. As deep learning architectures become more opaque, they risk becoming black boxes, hindering both scientific discovery and regulatory acceptance. For scientists, a lack of interpretability prevents the extraction of novel biological insights from the patterns a model has learned. For regulatory agencies, a clear design rationale is crucial for approval. The development of explainable AI (XAI) tailored for biomolecular science is therefore not an academic luxury but a practical necessity. By analyzing attention maps, learning feature attributions, and developing inherently interpretable model architectures, we can begin to understand why a model makes a certain prediction, facilitating the regulatory assessment and scientific validation of AI-designed therapeutics.

7.4. *Translational challenges and strategic considerations*

Finally, despite the significant potential of AI in the design of biologics and delivery systems, clinical translation of these innovations remains a substantial hurdle. Biologics and delivery systems are complex, heterogeneous, and highly sensitive to manufacturing processes. Even when AI-designed candidates demonstrate promising *in vitro* or preclinical performance, clinical translation still faces high attrition rates, reflecting the disconnect between early-stage design and clinical outcomes. Current AI workflows are often confined to early discovery and preclinical development, with limited integration of emerging clinical data, which constrains their ability to anticipate human-specific responses and guide later-stage optimization.

Regulatory compliance, scalability, and chemistry, manufacturing, and controls (CMC) are critical factors for clinical translation. Many AI-designed proteins, nucleic acids, or delivery systems face technical hurdles in synthesis, stabilization, and clinical-grade production, necessitating thorough evaluation of safety, efficacy, and long-term toxicity. By integrating multi-modal preclinical and clinical datasets, AI can support regulatory decision-making, guide dosing strategies, and predict drug performance, while integrating synthetic feasibility and manufacturing constraints directly into the design process.

Overall, while AI has the potential to accelerate biologic design and provide strategic guidance for clinical translation, realizing its full impact requires the integration of AI models across the entire development continuum—from early discovery through preclinical and clinical stages—to systematically address failure risks and improve clinical success rates.

8. Conclusions

Biologic drug development is undergoing a significant transformation driven by AI. As this review has detailed, AI has evolved from a peripheral tool for data analysis into a central, creative force, shifting macromolecular design from empirical screening to predictive engineering. By constructing and navigating complex fitness landscapes *in silico*, AI allows for the exploration of a design space far vaster than what is accessible through experimentation alone.

Future progress requires addressing the strategic frontiers outlined above. The future of the field will depend on the synergistic co-development of AI algorithms and high-throughput, automated experimental platforms to generate the high-quality, AI-native data needed to fuel them. Overcoming the profound complexity gap between molecular predictions and *in vivo* systemic behavior, while ensuring our models are both controllable and interpretable, will be essential for building trust and deriving true scientific insight.

In conclusion, the integration of artificial intelligence into biological macromolecular science is not an incremental improvement but a paradigm shift. We are moving toward a future where closed-loop, autonomous systems can design, synthesize,

and test therapeutic candidates with minimal human intervention. This evolution aims to accelerate the hit-to-lead phase and improve the precision of therapeutic discovery.

Acknowledgments

This research was supported in part by the National Natural Science Foundation of China (82425104 to Honglin Li), the National Key Research and Development Program of China (2022YFC3400501, 2022YFC3400504), and the Fundamental Research Funds for the Central Universities. Shiliang Li is also sponsored by the Shanghai Rising-Star Program (23QA1402800, China).

Author contributions

Jianxin Tang and Daohong Gong designed and wrote the paper and prepared figures. Honglin Li and Shiliang Li revised the manuscript. All of the authors have read and approved the final manuscript.

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Chandra P, Shahjad, Sharma KK, Verma A. Role of macromolecules in medical application: challenges and opportunities. *Macromol Symp* 2024;**413**:2300110.
2. Manivannan RK, Sharma N, Kumar V, Jayaraj I, Vimal S, Umesh M. A comprehensive review on natural macromolecular biopolymers for biomedical applications: recent advancements, current challenges, and future outlooks. *Carbohydr Polym Technol Appl* 2024;**8**:100536.
3. Liu W, Deng W, Hu L, Zou H. Advances in TRPV6 inhibitors for tumors by targeted therapies: macromolecular protein, synthetic small molecule compounds, and natural compounds. *Eur J Med Chem* 2024;**270**:116379.
4. Arora K, Sherilraj P, Abutwaibe K, Dhruw B, Mudavath SL. Exploring glycans as vital biological macromolecules: a comprehensive review of advancements in biomedical frontiers. *Int J Biol Macromol* 2024;**268**:131511.
5. Singh D. A short appraisal of biological macromolecules as tethering excipients for improved drug delivery: current advances. *J Macromol Sci Part B* 2024;**63**:604−20.
6. Mekala JR, Nalluri HP, Reddy PN, SB S, NS SK, GVSD SK, et al. Emerging trends and therapeutic applications of monoclonal antibodies. *Gene* 2024;**925**:148607.
7. Boroujeni MB, Mohebi S, Malekian A, Shahraeini SS, Gharagheizi Z, Shahkolahi S, et al. The therapeutic effect of engineered phage, derived protein and enzymes against superbug bacteria. *Biotechnol Bioeng* 2024;**121**:82−99.
8. Qiao Q, Wang J, Long K, Li L, Chen J, Guo Y, et al. A cascaded enzyme system based on the catalase-like activity of Ti3C2Tx MXene nanosheets for the efficient combination cancer therapy. *Nano Today* 2024;**54**:102059.
9. D'Alessio AM, Boffa I, De Stefano L, Soria LR, Brunetti-Pierri N. Liver gene transfer for metabolite detoxification in inherited metabolic diseases. *FEBS Lett* 2024;**598**:2372−84.
10. Ebenezer O, Oyebamiji AK, Olanlokun JO, Tuszynski JA, Wong GK-S. Recent update on siRNA therapeutics. *Int J Mol Sci* 2025;**26**:3456.
11. Parhiz H, Atochina-Vasserman EN, Weissman D. mRNA-based therapeutics: looking beyond COVID-19 vaccines. *Lancet* 2024;**403**:1192−204.
12. Liu F, Mailhot O, Glenn IS, Vigneron SF, Bassim V, Xu X, et al. The impact of library size and scale of testing on virtual screening. *Nat Chem Biol* 2025;**21**:1039−45.
13. Xue N, Hong D, Zhang D, Wang Q, Zhang S, Yang L, et al. Engineering IscB to develop highly efficient miniature editing tools in mammalian cells and embryos. *Mol Cell* 2024;**84**:3128−40.e4.
14. An M, Raguram A, Du SW, Banskota S, Davis JR, Newby GA, et al. Engineered virus-like particles for transient delivery of prime editor ribonucleoprotein complexes *in vivo*. *Nat Biotechnol* 2024;**42**:1526−37.
15. Raguram A, An M, Chen PZ, Liu DR. Directed evolution of engineered virus-like particles with improved production and transduction efficiencies. *Nat Biotechnol* 2025;**43**:1635−47.
16. Kumar N, Srivastava R. Deep learning in structural bioinformatics: current applications and future perspectives. *Briefings Bioinf* 2024;**25**:bbae042.
17. Hwang H, Jeon H, Yeo N, Baek D. Big data and deep learning for RNA biology. *Exp Mol Med* 2024;**56**:1293−321.
18. Jänes J, Beltrao P. Deep learning for protein structure prediction and design—progress and applications. *Mol Syst Biol* 2024;**20**:162−9.
19. Chakraborty C, Bhattacharya M, Pal S, Lee S-S. From machine learning to deep learning: advances of the recent data-driven paradigm shift in medicine and healthcare. *Curr Res Biotechnol* 2024;**7**:100164.
20. Diaz DJ, Gong C, Ouyang-Zhang J, Loy JM, Wells J, Yang D, et al. Stability oracle: a structure-based graph-transformer framework for identifying stabilizing mutations. *Nat Commun* 2024;**15**:6170.
21. Cheng P, Mao C, Tang J, Yang S, Cheng Y, Wang W, et al. Zero-shot prediction of mutation effects with multimodal deep representation learning guides protein engineering. *Cell Res* 2024;**34**:630−47.
22. Zhou Z, Zhang L, Yu Y, Wu B, Li M, Hong L, et al. Enhancing efficiency of protein language models with minimal wet-lab data through few-shot learning. *Nat Commun* 2024;**15**:5566.
23. Feng Z, Chen J, Hai Y, Pang X, Zheng K, Xie C, et al. Sliding-attention transformer neural architecture for predicting T cell receptor-antigen-human leucocyte antigen binding. *Nat Mach Intell* 2024;**6**:1216−30.
24. Garon EB, Rizvi NA, Hui R, Leighl N, Balmanoukian AS, Eder JP, et al. Pembrolizumab for the treatment of non-small-cell lung cancer. *N Engl J Med* 2015;**372**:2018−28.
25. Robert C, Schachter J, Long GV, Arance A, Grob JJ, Mortier L, et al. Pembrolizumab *versus* ipilimumab in advanced melanoma. *N Engl J Med* 2015;**372**:2521−32.
26. Wolchok JD, Kluger H, Callahan MK, Postow MA, Rizvi NA, Lesokhin AM, et al. Nivolumab plus ipilimumab in advanced melanoma. *N Engl J Med* 2013;**369**:122−33.
27. Borghaei H, Paz-Ares L, Horn L, Spigel DR, Steins M, Ready NE, et al. Nivolumab *versus* docetaxel in advanced nonsquamous non-small-cell lung cancer. *N Engl J Med* 2015;**373**:1627−39.
28. Weinblatt ME, Keystone EC, Furst DE, Moreland LW, Weisman MH, Birbara CA, et al. Adalimumab, a fully human anti-tumor necrosis factor α monoclonal antibody, for the treatment of rheumatoid arthritis in patients taking concomitant methotrexate: the ARMADA trial. *Arthritis Rheum* 2003;**48**:35−45.
29. Griffiths CE, Strober BE, Van De Kerkhof P, Ho V, Fidelus-Gort R, Yeilding N, et al. Comparison of ustekinumab and etanercept for moderate-to-severe psoriasis. *N Engl J Med* 2010;**362**:118−28.
30. Simoes EA, Groothuis JR, Carbonell-Estrany X, Rieger CH, Mitchell I, Fredrick LM, et al. Palivizumab prophylaxis, respiratory syncytial virus, and subsequent recurrent wheezing. *J Pediatr* 2007;**151**:34−42.
31. Goadsby PJ, Reuter U, Hallström Y, Broessner G, Bonner JH, Zhang F, et al. A controlled trial of erenumab for episodic migraine. *N Engl J Med* 2017;**377**:2123−32.

32. Marso SP, Daniels GH, Brown-Frandsen K, Kristensen P, Mann JF, Nauck MA, et al. Liraglutide and cardiovascular outcomes in type 2 diabetes. *N Engl J Med* 2016;**375**:311−22.
33. Marso SP, Bain SC, Consoli A, Eliaschewitz FG, Jódar E, Leiter LA, et al. Semaglutide and cardiovascular outcomes in patients with type 2 diabetes. *N Engl J Med* 2016;**375**:1834−44.
34. Saag KG, Shane E, Boonen S, Marín F, Donley DW, Taylor KA, et al. Teriparatide or alendronate in glucocorticoid-induced osteoporosis. *N Engl J Med* 2007;**357**:2028−39.
35. Crawford ED, Eisenberger MA, McLeod DG, Spaulding JT, Benson R, Dorr FA, et al. A controlled trial of leuprolide with and without flutamide in prostatic carcinoma. *N Engl J Med* 1989;**321**: 419−24.
36. Cooper BM, Iegre J, O'Donovan DH, Halvarsson MÖ, Spring DR. Peptides as a platform for targeted therapeutics for cancer: peptide−drug conjugates (PDCs). *Chem Soc Rev* 2021;**50**:1480−94.
37. Velkov T, Thompson PE, Nation RL, Li J. Structure−activity relationships of polymyxin antibiotics. *J Med Chem* 2010;**53**: 1898−916.
38. Dodick DW, Lipton RB, Ailani J, Lu K, Finnegan M, Trugman JM, et al. Ubrogepant for the treatment of migraine. *N Engl J Med* 2019; **381**:2230−41.
39. Pastores GM. Recombinant glucocerebrosidase (imiglucerase) as a therapy for Gaucher disease. *BioDrugs* 2010;**24**:41−7.
40. Bolli GB, Owens DR. Insulin glargine. *Lancet* 2000;**356**:443−5.
41. Deal CL, Steelman J, Vlachopapadopoulou E, Stawerska R, Silverman LA, Phillip M, et al. Efficacy and safety of weekly somatrogon *vs* daily somatropin in children with growth hormone deficiency: a phase 3 study. *J Clin Endocrinol Metab* 2022;**107**:e2717−28.
42. Klamroth R, Simpson M, von Depka-Prondzinski M, Gill J, Morfini M, Powell J, et al. Comparative pharmacokinetics of rVIII-SingleChain and octocog alfa (Advate®) in patients with severe haemophilia A. *Haemophilia* 2016;**22**:730−8.
43. Singh AK, Szczech L, Tang KL, Barnhart H, Sapp S, Wolfson M, et al. Correction of anemia with epoetin alfa in chronic kidney disease. *N Engl J Med* 2006;**355**:2085−98.
44. Dale DC, Crawford J, Klippel Z, Reiner M, Osslund T, Fan E, et al. A systematic literature review of the efficacy, effectiveness, and safety of filgrastim. *Support Care Cancer* 2018;**26**:7−20.
45. Wee Yong V, Chabot S, Stuve O, Williams G. Interferon beta in the treatment of multiple sclerosis: mechanisms of action. *Neurology* 1998;**51**:682−9.
46. Mercuri E, Darras BT, Chiriboga CA, Day JW, Campbell C, Connolly AM, et al. Nusinersen *versus* sham control in later-onset spinal muscular atrophy. *N Engl J Med* 2018;**378**:625−35.
47. Adams D, Gonzalez-Duarte A, O'Riordan WD, Yang C-C, Ueda M, Kristen AV, et al. Patisiran, an RNAi therapeutic, for hereditary transthyretin amyloidosis. *N Engl J Med* 2018;**379**:11−21.
48. Frangoul H, Altshuler D, Cappellini MD, Chen Y-S, Domm J, Eustace BK, et al. CRISPR-Cas9 gene editing for sickle cell disease and β-thalassemia. *N Engl J Med* 2021;**384**:252−60.
49. Ng EW, Shima DT, Calias P, Cunningham ET Jr, Guyer DR, Adamis AP. Pegaptanib, a targeted anti-VEGF aptamer for ocular vascular disease. *Nat Rev Drug Discov* 2006;**5**:123−32.
50. Baden LR, El Sahly HM, Essink B, Kotloff K, Frey S, Novak R, et al. Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *N Engl J Med* 2021;**384**:403−16.
51. Polack FP, Thomas SJ, Kitchin N, Absalon J, Gurtman A, Lockhart S, et al. Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *N Engl J Med* 2020;**383**:2603−15.
52. Ogbonmide T, Rathore R, Rangrej SB, Hutchinson S, Lewis M, Ojilere S, et al. Gene therapy for spinal muscular atrophy (SMA): a review of current challenges and safety considerations for Onasemnogene Abeparvovec (Zolgensma). *Cureus* 2023;**15**:e36197.
53. Prado DA, Acosta-Acero M, Maldonado RS. Gene therapy beyond luxturna: a new horizon of the treatment for inherited retinal disease. *Curr Opin Ophthalmol* 2020;**31**:147−54.
54. Awasthi R, Maier HJ, Zhang J, Lim S. Kymriah®(tisagenlecleucel)—An overview of the clinical development journey of the first approved CAR-T therapy. *Hum Vaccines Immunother* 2023;**19**:2210046.
55. Papadouli I, Mueller-Berghaus J, Beuneu C, Ali S, Hofner B, Petavy F, et al. EMA review of axicabtagene ciloleucel (Yescarta) for the treatment of diffuse large B-cell lymphoma. *Oncologist* 2020;**25**: 894−902.
56. Verma S, Miles D, Gianni L, Krop IE, Welslau M, Baselga J, et al. Trastuzumab emtansine for HER2-positive advanced breast cancer. *N Engl J Med* 2012;**367**:1783−91.
57. Hennrich U, Kopka K. Lutathera®: the first FDA-and EMA-approved radiopharmaceutical for peptide receptor radionuclide therapy. *Pharmaceuticals* 2019;**12**:114.
58. Ray KK, Wright RS, Kallend D, Koenig W, Leiter LA, Raal FJ, et al. Two phase 3 trials of inclisiran in patients with elevated LDL cholesterol. *N Engl J Med* 2020;**382**:1507−19.
59. Farghali H, Kutinová Canová N, Arora M. The potential applications of artificial intelligence in drug discovery and development. *Physiol Res* 2021:S715−22.
60. Arnold C. Inside the nascent industry of AI-designed drugs. *Nat Med* 2023;**29**:1292−5.
61. Burki T. A new paradigm for drug development. *Lancet Digit Health* 2020;**2**:e226−7.
62. Payne A, Fons P, Alt I, van Ham J, Taubert C, Bell A, et al. Abstract 1731: EXS21546, a non-CNS penetrant A2AR-selective antagonist for anti-cancer immunotherapy. *Cancer Res* 2021;**81**:1731.
63. Ren F, Aliper A, Chen J, Zhao H, Rao S, Kuppe C, et al. A small-molecule TNIK inhibitor targets fibrosis in preclinical and clinical models. *Nat Biotechnol* 2025;**43**:63−75.
64. Xu Z, Ren F, Wang P, Cao J, Tan C, Ma D, et al. A generative AI-discovered TNIK inhibitor for idiopathic pulmonary fibrosis: a randomized phase 2a trial. *Nat Med* 2025;**31**:2602−10.
65. Mørk SK, Kadivar M, Bol KF, Draghi A, Westergaard MCW, Skadborg SK, et al. Personalized therapy with peptide-based neoantigen vaccine (EVX-01) including a novel adjuvant, CAF®09b, in patients with metastatic melanoma. *OncoImmunology* 2022;**11**:2023255.
66. Perez SM, Dimastromatteo J, Landen CN, Kelly KA. A novel monoclonal antibody targeting cancer-specific plectin has potent antitumor activity in ovarian cancer. *Cells* 2021;**10**:2218.
67. Kjølbye LR, Pereira GP, Bartocci A, Pannuzzo M, Albani S, Marchetto A, et al. Towards design of drugs and delivery systems with the Martini coarse-grained model. *QRB Discov* 2022;**3**:e19.
68. Huang D, Pachuda N, Sauer JM, Dobbins D, Steckbeck J. The engineered antibiotic peptide PLG0206 eliminates biofilms and is a potential treatment for periprosthetic joint infections. *Antibiotics* 2022;**11**:41.
69. Singh P, Toom S, Huang Y. Anti-claudin 18.2 antibody as new targeted therapy for advanced gastric cancer. *J Hematol Oncol* 2017;**10**:105.
70. Nagra NS, van der Veken L, Stanzl E, Champagne D, Devereson A, Macak M. The company landscape for artificial intelligence in large-molecule drug discovery. *Nat Rev Drug Discov* 2023;**22**:949−50.
71. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inf Assoc* 2011;**18**:544−51.
72. Meier J, Rao R, Verkuil R, Liu J, Sercu T, Rives A. Language models enable zero-shot prediction of the effects of mutations on protein function. *Adv Neural Inf Process Syst* 2021;**34**:29287−303.
73. The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res* 2015;**43**:D204−12.
74. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. *Nucleic Acids Res* 2013;**41**:D36−42.
75. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. *arXiv preprint* 2019. 1907.11692.
76. Singh J, Paliwal K, Litfin T, Singh J, Zhou Y. Reaching alignment-profile-based accuracy in predicting protein secondary and tertiary structural properties without alignment. *Sci Rep* 2022;**12**:7607.

77. Zhang X, Tseo Y, Bai Y, Chen F, Uhler C. Prediction of protein subcellular localization in single cells. *Nat Methods* 2025;**22**: 1265−75.
78. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;**379**:1123−30.
79. Jiang TT, Fang L, Wang K. Deciphering "the language of nature": a transformer-based language model for deleterious mutations in proteins. *Innovation* 2023;**4**:100487.
80. Nijkamp E, Ruffolo JA, Weinstein EN, Naik N, Madani A. Progen2: exploring the boundaries of protein language models. *Cell Syst* 2023; **14**:968−78.
81. Madani A, Krause B, Greene ER, Subramanian S, Mohr BP, Holton JM, et al. Large language models generate functional protein sequences across diverse families. *Nat Biotechnol* 2023;**41**: 1099−106.
82. Bhatnagar A, Jain S, Beazer J, Curran SC, Hoffnagle AM, Ching K, et al. Scaling unlocks broader generation and deeper functional understanding of proteins. *bioRxiv* 2025. 2025.04.15.649055.
83. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. *OpenAI Techn Rep* 2018. Available from: https://gwern.net/doc/www/s3-us-west-2.amazonaws.com/d73fdc5ffa8627bce44dcda2fc012da638ffb158.pdf.
84. Hayes T, Rao R, Akin H, Sofroniew NJ, Oktay D, Lin Z, et al. Simulating 500 million years of evolution with a language model. *Science* 2025;**387**:850−8.
85. Su J, Han C, Zhou Y, Shan J, Zhou X, Yuan F. SaProt: protein language modeling with structure-aware vocabulary. *bioRxiv* 2023. 2023.10.01.560349.
86. Liu S, Nie W, Wang C, Lu J, Qiao Z, Liu L, et al. Multi-modal molecule structure-text model for text-based retrieval and editing. *Nat Mach Intell* 2023;**5**:1447−57.
87. Xiao H, Lin W, Chen X, Wang H, Chen K, Li J, et al. STELLA: towards protein function prediction with multimodal LLMs integrating sequence-structure representations. *arXiv preprint* 2025. 2506.03800.
88. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, et al. Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* 2023;**42**:243−6.
89. Sanabria M, Hirsch J, Joubert PM, Poetsch AR. DNA language model GROVER learns sequence context in the human genome. *Nat Mach Intell* 2024;**6**:911−23.
90. Shen T, Hu Z, Sun S, Liu D, Wong F, Wang J, et al. Accurate RNA 3D structure prediction using a language model-based deep learning approach. *Nat Methods* 2024;**21**:2287−98.
91. Wang N, Bian J, Li Y, Li X, Mumtaz S, Kong L, et al. Multi-purpose RNA language modelling with motif-aware pretraining and type-guided fine-tuning. *Nat Mach Intell* 2024;**6**:548−57.
92. Ji Y, Zhou Z, Liu H, Davuluri RV. DNABERT: pre-trained bidirectional encoder representations from Transformers model for DNA-language in genome. *Bioinformatics* 2021;**37**:2112−20.
93. Nguyen E, Poli M, Durrant MG, Kang B, Katrekar D, Li DB, et al. Sequence modeling and design from molecular to genome scale with Evo. *Science* 2024;**386**:eado9336.
94. Kumar H, Kim P. Artificial intelligence in fusion protein three-dimensional structure prediction: review and perspective. *Clin Transl Med* 2024;**14**:e1789.
95. Vincoff S, Goel S, Kholina K, Pulugurta R, Vure P, Chatterjee P. FusOn-pLM: a fusion oncoprotein-specific language model *via* adjusted rate masking. *Nat Commun* 2025;**16**:1436.
96. Tripathi S, Shirnekhi HK, Gorman SD, Chandra B, Baggett DW, Park C-G, et al. Defining the condensate landscape of fusion oncoproteins. *Nat Commun* 2023;**14**:6008.
97. Rohl CA, Strauss CEM, Misura KMS, Baker D. Protein structure prediction using Rosetta. *Methods Enzymol* 2004;**383**:66−93.
98. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583−9.
99. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 2021;**373**: 871−6.
100. Yang T, Han M, Wen X, Zheng Y. Advances in intelligent design and optimization methods for nucleic acid sequences. *J Artif Intell Bioinforma* 2025;**1**:12−29.
101. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 2024;**630**:493−500.
102. Krishna R, Wang J, Ahern W, Sturmfels P, Venkatesh P, Kalvet I, et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* 2024;**384**:eadl2528.
103. Lewis S, Hempel T, Jiménez-Luna J, Gastegger M, Xie Y, Foong AY, et al. Scalable emulation of protein equilibrium ensembles with generative deep learning. *bioRxiv* 2024. 2024.12.05.626885.
104. Leman JK, Weitzner BD, Lewis SM, Adolf-Bryfogle J, Alam N, Alford RF, et al. Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat Methods* 2020;**17**:665−80.
105. Liu Y, Kuhlman B. RosettaDesign server for protein design. *Nucleic Acids Res* 2006;**34**:W235−8.
106. Song J, Meng C, Ermon S. Denoising diffusion implicit models. *arXiv preprint* 2020. 2010.02502.
107. Geffner T, Didi K, Zhang Z, Reidenbach D, Cao Z, Yim J, et al. Proteina: scaling flow-based protein structure generative models. *arXiv preprint* 2025. 2503.00710.
108. Chan SH. Tutorial on diffusion models for imaging and vision. *arXiv preprint* 2024. 2403.18103.
109. Watson JL, Juergens D, Bennett NR, Trippe BL, Yim J, Eisenach HE, et al. *De novo* design of protein structure and function with RFdiffusion. *Nature* 2023;**620**:1089−100.
110. Bai P, Miljković F, Liu X, De Maria L, Croasdale-Wood R, Rackham O, et al. Mask-prior-guided denoising diffusion improves inverse protein folding. *Nat Mach Intell* 2025:1−13.
111. Dauparas J, Lee GR, Pecoraro R, An L, Anishchenko I, Glasscock C, et al. Atomic context-conditioned protein sequence design using LigandMPNN. *Nat Methods* 2025;**22**:717−23.
112. Zhou B, Zheng L, Wu B, Yi K, Zhong B, Tan Y, et al. A conditional protein diffusion model generates artificial programmable endonuclease sequences with enhanced activity. *Cell Discov* 2024;**10**:1−12.
113. Gao Z, Tan C, Li SZ. PiFold: toward effective and efficient protein inverse folding. *arXiv preprint* 2022. 2209.12643.
114. Hsu C, Verkuil R, Liu J, Lin Z, Hie B, Sercu T, et al. Learning inverse folding from millions of predicted structures. *bioRxiv* 2022. 2022.04.10.487779.
115. Zheng Z, Deng Y, Xue D, Zhou Y, Ye F, Gu Q. Structure-informed language models are protein designers. *arXiv preprint* 2023. 2302.01649.
116. Dauparas J, Anishchenko I, Bennett N, Bai H, Ragotte RJ, Milles LF, et al. Robust deep learning-based protein sequence design using ProteinMPNN. *Science* 2022;**378**:49−56.
117. Ingraham JB, Baranov M, Costello Z, Barber KW, Wang W, Ismail A, et al. Illuminating protein space with a programmable generative model. *Nature* 2023;**623**:1070−8.
118. Frank C, Khoshouei A, Fuβ L, Schiwietz D, Putz D, Weber L, et al. Scalable protein design using optimization in a relaxed sequence space. *Science* 2024;**386**:439−45.
119. Meshchaninov V, Strashnov P, Shevtsov A, Nikolaev F, Ivanisenko N, Kardymon O, et al. Diffusion on language model encodings for protein sequence generation. *arXiv preprint* 2025;**2403**:03726.
120. Anand R, Joshi CK, Morehead A, Jamasb AR, Harris C, Mathis SV, et al. RNA-FrameFlow: flow matching for *de novo* 3D RNA backbone design. *arXiv preprint* 2025;**2406**:13839.

121. Ma R, Zhang Z, Wang Z, Hua C, Zhou Z, Cao F, et al. RiboFlow: conditional *de novo* RNA sequence-structure co-design *via* synergistic flow matching. *arXiv preprint* 2025;**2503**:17007.
122. Hou D, Zhang S, Ma M, Lin H, Wan Z, Zhao H, et al. A hyperbolic discrete diffusion 3D RNA inverse folding model for functional RNA design. *J Chem Inf Model* 2025;**65**:6568−84.
123. Wong F, He D, Krishnan A, Hong L, Wang AZ, Wang J, et al. Deep generative design of RNA aptamers using structural predictions. *Nat Comput Sci* 2024;**4**:829−39.
124. Berger S, Seeger F, Yu T-Y, Aydin M, Yang H, Rosenblum D, et al. Preclinical proof of principle for orally delivered Th17 antagonist miniproteins. *Cell* 2024;**187**:4305−17.e18.
125. Adams CS, Kim H, Burtner AE, Lee DS, Dobbins C, Criswell C, et al. *De novo* design of protein minibinder agonists of TLR3. *Nat Commun* 2025;**16**:1234.
126. Chu AE, Lu T, Huang P-S. Sparks of function by *de novo* protein design. *Nat Biotechnol* 2024;**42**:203−15.
127. Bennett NR, Coventry B, Goreshnik I, Huang B, Allen A, Vafeados D, et al. Improving *de novo* protein binder design with deep learning. *Nat Commun* 2023;**14**:2625.
128. Vázquez Torres S, Benard Valle M, Mackessy SP, Menzies SK, Casewell NR, Ahmadi S, et al. *De novo* designed proteins neutralize lethal snake venom toxins. *Nature* 2025;**639**:225−31.
129. Pacesa M, Nickel L, Schellhaas C, Schmidt J, Pyatova E, Kissling L, et al. One-shot design of functional protein binders with BindCraft. *Nature* 2025;**646**:483−92.
130. Korendovych IV, DeGrado WF. *De novo* protein design, a retrospective. *Q Rev Biophys* 2020;**53**:e3.
131. Jelokhani-Niaraki M. Membrane proteins: structure, function and motion. *Int J Mol Sci* 2023;**24**:468.
132. Hauser AS, Attwood MM, Rask-Andersen M, Schiöth HB, Gloriam DE. Trends in GPCR drug discovery: new agents, targets and indications. *Nat Rev Drug Discov* 2017;**16**:829−42.
133. Congreve M, de Graaf C, Swain NA, Tate CG. Impact of GPCR structures on drug discovery. *Cell* 2020;**181**:81−91.
134. Muratspahić E, Feldman D, Kim DE, Qu X, Bratovianu A-M, Rivera-Sánchez P, et al. *De novo* design of miniprotein agonists and antagonists targeting G protein-coupled receptors. *bioRxiv* 2025. 2025.03.23.644666.
135. Goverde CA, Pacesa M, Goldbach N, Dornfeld LJ, Balbi PEM, Georgeon S, et al. Computational design of soluble and functional membrane protein analogues. *Nature* 2024;**631**:449−58.
136. Goel S, Schray PM, Zhang Y, Vincoff S, Kratochvil HT, Chatterjee P. Token-level guided discrete diffusion for membrane protein design. *arXiv preprint* 2025. 2410.16735.
137. Vorobieva AA, White P, Liang B, Horne JE, Bera AK, Chow CM, et al. *De novo* design of transmembrane β barrels. *Science* 2021;**371**:eabc8182.
138. He H, He B, Guan L, Zhao Y, Jiang F, Chen G, et al. *De novo* generation of SARS-CoV-2 antibody CDRH3 with a pre-trained generative large language model. *Nat Commun* 2024; **15**:6867.
139. Guloglu B, Bragança M, Graves A, Cameron S, Atkinson T, Copoiu L, et al. AbBFN2: a flexible antibody foundation model based on Bayesian flow networks. *bioRxiv* 2025. 2025.04.29.651170.
140. Bennett NR, Watson JL, Ragotte RJ, Borst AJ, See DL, Weidle C, et al. Atomically accurate *de novo* design of antibodies with RFdiffusion. *bioRxiv* 2025. 2024.03.14.585103.
141. Wang R, Wu F, Gao X, Wu J, Zhao P, Yao J. IgGM: a generative model for functional antibody and nanobody design. *bioRxiv* 2024. 2024.09.19.613838.
142. Rettie SA, Campbell KV, Bera AK, Kang A, Kozlov S, Bueso YF, et al. Cyclic peptide structure prediction and design using AlphaFold2. *Nat Commun* 2025;**16**:4730.
143. Diao Y, Liu D, Ge H, Zhang R, Jiang K, Bao R, et al. Macrocyclization of linear molecules by deep learning to facilitate macrocyclic drug candidates discovery. *Nat Commun* 2023;**14**:4552.
144. Bhat S, Palepu K, Hong L, Mao J, Ye T, Iyer R, et al. *De novo* design of peptide binders to conformationally diverse targets with contrastive language modeling. *Sci Adv* 2025;**11**:eadr8638.
145. Wang J, Feng J, Kang Y, Pan P, Ge J, Wang Y, et al. Discovery of antimicrobial peptides with notable antibacterial potency by an LLM-based foundation model. *Sci Adv* 2025;**11**. eads8932.
146. Feng Z, Pang X, Xu Q, Gu Z, Li S, Zhu L, et al. Design of a multivalent Multi-valent SARS-CoV-2 peptide vaccine for broad immune protection broad immune protection *via* deep learning. *Engineering* 2025;**52**:142−59.
147. Geylan G, Paul Janet J, Tibo A, He J, Patronov A, Kabeshov M, et al. PepINVENT: generative peptide design beyond natural amino acids. *Chem Sci* 2025;**16**:8682−96.
148. Vázquez Torres S, Leung PJY, Venkatesh P, Lutz ID, Hink F, Huynh H-H, et al. *De novo* design of high-affinity binders of bioactive helical peptides. *Nature* 2024;**626**:435−42.
149. Pokorna S, Khersonsky O, Lipsh-Sokolik R, Goldenzweig A, Nielsen R, Ashani Y, et al. Design of a stable human acid-β-glucosidase: towards improved Gaucher disease therapy and mutation classification. *FEBS J* 2023;**290**:3383−99.
150. Hettiaratchi MH, O'Meara MJ, O'Meara TR, Pickering AJ, Letko-Khait N, Shoichet MS. Reengineering biocatalysts: computational redesign of chondroitinase ABC improves efficacy and stability. *Sci Adv* 2020;**6**:eabc6378.
151. Chen G, Zhao K, Li M, Zhang Y, Li T, Yuan J, et al. Rapid clearance of achiral small-molecule drugs using *de novo*-designed proteins and their cyclic and mirror-image variants. *Nat Biomed Eng* 2025;**9**: 1775−87.
152. Ruffolo JA, Nayfach S, Gallagher J, Bhatnagar A, Beazer J, Hussain R, et al. Design of highly functional genome editors by modelling CRISPR-Cas sequences. *Nature* 2025;**645**:518−25.
153. Nammi B, Jayasinghe-Arachchige VM, Madugula SS, Artiles M, Radler CN, Pham T, et al. CasGen: a regularized generative model for CRISPR cas protein design with classification and margin-based optimization. *bioRxiv* 2025. 2025.02.28.640911.
154. Park J-C, Uhm H, Kim Y-W, Oh YE, Lee JH, Yang J, et al. AI-generated MLH1 small binder improves prime editing efficiency. *Cell* 2025;**188**:5831−46.e21.
155. Tang Q, Khvorova A. RNAi-based drug design: considerations and future directions. *Nat Rev Drug Discov* 2024;**23**:341−64.
156. Li Y, Garcia Jr G, Arumugaswami V, Guo F. Structure-based design of antisense oligonucleotides that inhibit SARS-CoV-2 replication. *bioRxiv* 2021. 2021.08.23.457434.
157. Guo G, Guo L, Qian J, He X, Qian X, Wang L, et al. *De novo* design of protein-binding aptamers through deep reinforcement learning assembly of nucleic acid fragments. *bioRxiv* 2025. 2025.06.01.657174.
158. Zhang H, Liu H, Xu Y, Liu Y, Wang J, Qin Y, et al. Deep generative models generate mRNA sequences with enhanced translation capacity and stability. *bioRxiv* 2024. 2024.06.20.599727.
159. Tang X, Huo M, Chen Y, Huang H, Qin S, Luo J, et al. A novel deep generative model for mRNA vaccine development: designing 5′ UTRs with *N*1-methyl-pseudouridine modification. *Acta Pharm Sin B* 2024;**14**:1814−26.
160. Zhao Y, Oono K, Takizawa H, Kotera M. GenerRNA: a generative pre-trained language model for *de novo* RNA design. *PLoS One* 2024;**19**:e0310814.
161. Morrow AK, Thornal A, Flynn ED, Hoelzli E, Shan M, Garipler G, et al. ML-driven design of 3′UTRs for mRNA stability. *bioRxiv* 2024. 2024.10.07.616676.
162. Sciabola S, Xi H, Cruz D, Cao Q, Lawrence C, Zhang T, et al. PFRED: a computational platform for siRNA and antisense oligonucleotides design. *PLoS One* 2021;**16**:e0238753.
163. Chiba S, Lim KRQ, Sheri N, Anwar S, Erkut E, Shah MNA, et al. eSkip-Finder: a machine learning-based web application and database to identify the optimal sequences of antisense oligonucleotides for exon skipping. *Nucleic Acids Res* 2021;**49**:W193−8.

164. Hörberg J, Carlesso A, Reymer A. Mechanistic insights into ASO-RNA complexation: advancing antisense oligonucleotide design strategies. *Mol Ther Nucleic Acids* 2024;**35**:102351.
165. Andress C, Kappel K, Villena ME, Cuperlovic-Culf M, Yan H, Li Y. DAPTEV: deep aptamer evolutionary modelling for COVID-19 drug design. *PLoS Comput Biol* 2023;**19**:e1010774.
166. Baisya D, Ramesh A, Schwartz C, Lonardi S, Wheeldon I. Genome-wide functional screens enable the prediction of high activity CRISPR-Cas9 and -Cas12a guides in Yarrowia lipolytica. *Nat Commun* 2022;**13**:922.
167. Louie W, Shen MW, Tahiry Z, Zhang S, Worstell D, Cassa CA, et al. Machine learning based CRISPR gRNA design for therapeutic exon skipping. *PLoS Comput Biol* 2021;**17**:e1008605.
168. Li C, Zhang R, Wang J, Wilson LM, Yan Y. Protein engineering for improving and diversifying natural product biosynthesis. *Trends Biotechnol* 2020;**38**:729−44.
169. Esvelt KM, Carlson JC, Liu DR. A system for the continuous directed evolution of biomolecules. *Nature* 2011;**472**:499−503.
170. Yang KK, Wu Z, Arnold FH. Machine-learning-guided directed evolution for protein engineering. *Nat Methods* 2019;**16**:687−94.
171. Pokharel S, Pratyush P, Heinzinger M, Newman RH, Kc DB. Improving protein succinylation sites prediction using embeddings from protein language model. *Sci Rep* 2022;**12**:16933.
172. Yang J, Lal RG, Bowden JC, Astudillo R, Hameedi MA, Kaur S, et al. Active learning-assisted directed evolution. *Nat Commun* 2025; **16**:714.
173. Emami P, Perreault A, Law J, Biagioni D, St, John P. Plug & play directed evolution of proteins with gradient-based discrete MCMC. *Mach Learn Sci Technol* 2023;**4**:025014.
174. Wang Y, Tang H, Huang L, Pan L, Yang L, Yang H, et al. Self-play reinforcement learning guides protein engineering. *Nat Mach Intell* 2023;**5**:845−60.
175. Jiang K, Yan Z, Di Bernardo M, Sgrizzi SR, Villiger L, Kayabolen A, et al. Rapid *in silico* directed evolution by a protein language model with EVOLVEpro. *Science* 2024;**387**:eadr6006.
176. Yang Y, Liu S, Luo Y, Wang B, Wang J, Li J, et al. High-throughput saturation mutagenesis generates a high-affinity antibody against SARS-CoV-2 variants using protein surface display assay on a human cell. *PLoS Pathog* 2023;**19**:e1011119.
177. Lee M, Vecchietti LF, Jung H, Ro HJ, Cha M, Kim HM. Robust optimization in protein fitness landscapes using reinforcement learning in latent space. *arXiv preprint* 2024;**2405**:18986.
178. Guan C, Wan F, Torres MDT, Fuente-Nunez C de la. Improving functional protein generation *via* foundation model-derived latent space likelihood optimization. *bioRxiv* 2025. 2025.01.07.631724.
179. Kim SJ, Park Y, Hong HJ. Antibody engineering for the development of therapeutic antibodies. *Mol Cells* 2005;**20**:17−29.
180. Svilenov HL, Sacherl J, Protzer U, Zacharias M, Buchner J. Mechanistic principles of an ultra-long bovine CDR reveal strategies for antibody design. *Nat Commun* 2021;**12**:6737.
181. Shanker VR, Bruun TUJ, Hie BL, Kim PS. Unsupervised evolution of protein and antibody complexes with a structure-informed language model. *Science* 2024;**385**:46−53.
182. Zeng Y, Elliott H, Maffettone P, Greenside P, Bastani O, Gardner JR. Antibody design with constrained bayesian optimization. In: *ICLR 2024 workshop on generative and experimental perspectives for biomolecular design*; 2024.
183. Wang L, Wang N, Zhang W, Cheng X, Yan Z, Shao G, et al. Therapeutic peptides: current applications and future directions. *Signal Transduct Targeted Ther* 2022;**7**:48.
184. Bryant P, Elofsson A. EvoBind: *in silico* directed evolution of peptide binders with AlphaFold. *bioRxiv* 2022. 2022.07.23.501214.
185. Chen LT, Quinn Z, Dumas M, Peng C, Hong L, Lopez-Gonzalez M, et al. Target sequence-conditioned design of peptide binders using masked language modeling. *Nat Biotechnol* 2026;**44**:1002−10.
186. Torres MDT, Chen LT, Wan F, Chatterjee P, Fuente-Nunez C de la. Generative latent diffusion language modeling yields anti-infective synthetic peptides. *Cell Biomater* 2025;**1**:100183.
187. Zhang R, Wu H, Xiu Y, Li K, Chen N, Wang Y, et al. PepLand: a large-scale pre-trained peptide representation model for a comprehensive landscape of both canonical and non-canonical amino acids. *arXiv preprint* 2023. 2311.04419.
188. Bae D, Kim M, Seo J, Nam H. AI-guided discovery and optimization of antimicrobial peptides through species-aware language model. *bioRxiv* 2025. 2025.05.20.654992.
189. Li H, Yang Y, Hong W, Huang M, Wu M, Zhao X. Applications of genome editing technology in the targeted therapy of human diseases: mechanisms, advances and prospects. *Signal Transduct Targeted Ther* 2020;**5**:1.
190. Hu J, Guo M, Gao Q, Jia H, He M, Wang Z, et al. QBEmax is a sequence-permuted and internally protected base editor. *Nat Biotechnol* 2026;**44**:387−93.
191. Chen Q, Sun Y, Yao J, Lu Y, Qiu R, Zhou F, et al. Engineering of peptide-inserted base editors with enhanced accuracy and security. *Small* 2025;**21**:2411583.
192. He Y, Zhou X, Chang C, Chen G, Liu W, Li G, et al. Protein language models-assisted optimization of a uracil-*N*-glycosylase variant enables programmable T-to-G and T-to-C base editing. *Mol Cell* 2024; **84**:1257−70.e6.
193. Kalemati M, Darvishi S, Koohi S. CapsNet-MHC predicts peptide-MHC class I binding based on capsule neural networks. *Commun Biol* 2023;**6**:492.
194. Wang H, Xia C, Zhu Y, Wang Y, Chu X, Pan X, et al. AI-guided precision in antibody humanization: structural modeling to minimize immunogenicity and preserve efficacy. *bioRxiv* 2025. 2025.03.09.641110.
195. Reeves S, Kalyaanamoorthy S. Zero-shot transfer of protein sequence likelihood models to thermostability prediction. *Nat Mach Intell* 2024;**6**:1063−76.
196. Cuturello F, Celoria M, Ansuini A, Cazzaniga A. Enhancing predictions of protein stability changes induced by single mutations using MSA-based language models. *Bioinformatics* 2024;**40**. btae447.
197. Dieckhaus H, Brocidiacono M, Randolph NZ, Kuhlman B. Transfer learning to leverage larger datasets for improved prediction of protein stability changes. *Proc Natl Acad Sci* 2024;**121**: e2314853121.
198. Zhang H, Zhang L, Lin A, Xu C, Li Z, Liu K, et al. Algorithm for optimized mRNA design improves stability and immunogenicity. *Nature* 2023;**621**:396−403.
199. Mo O, Zhang Z, Cheng X, Zhu L, Zhang K, Zhang N, et al. mRNAdesigner: an integrated web server for optimizing mRNA design and protein translation in eukaryotes. *Nucleic Acids Res* 2025; **53**:W415−26.
200. Outeiral C, Deane CM. Codon language embeddings provide strong signals for use in protein engineering. *Nat Mach Intell* 2024;**6**: 170−9.
201. Tang S, Zhang Y, Chatterjee P. PepTune: de novo generation of therapeutic peptides with multi-objective-guided discrete diffusion. *arXiv preprint* 2024. 2412.17780.
202. Kang S, Lee D, Hwang G, Lee K, Kang M. ASOptimizer: optimizing chemical diversity of antisense oligonucleotides through deep learning. *Nucleic Acids Res* 2025:gkaf392.
203. Xu Y, Ma S, Cui H, Chen J, Xu S, Gong F, et al. AGILE platform: a deep learning powered approach to accelerate LNP development for mRNA delivery. *Nat Commun* 2024;**15**:6305.
204. Huang J, Lin Q, Fei H, He Z, Xu H, Li Y, et al. Discovery of deaminase functions by structure-based protein clustering. *Cell* 2023; **186**:3182−95.e14.
205. Yoon PH, Zhang Z, Loi KJ, Adler BA, Lahiri A, Vohra K, et al. Structure-guided discovery of ancestral CRISPR-Cas13 ribonucleases. *Science* 2024;**385**:538−43.
206. Wolf B, Shehu P, Brenker L, von Bachmann A-L, Kroell A-S, Southern N, et al. Rational engineering of allosteric protein switches by *in silico* prediction of domain insertion sites. *Nat Methods* 2025; **22**:1698−706.

207. Plaper T, Merljak E, Fink T, Satler T, Ljubetič A, Lainšček D, et al. Designed allosteric protein logic. *Cell Discov* 2024;**10**:1−15.
208. Wang J, Ouyang X, Meng S, Zhao B, Liu L, Li C, et al. Rational multienzyme architecture design with iMARS. *Cell* 2025;**188**: 1349−62.e17.
209. Chen L, Li B, Chen Y, Lin M, Zhang S, Li C, et al. ADCNet: a unified framework for predicting the activity of antibody−drug conjugates. *arXiv preprint* 2024. 2401.09176.
210. Zheng J, Wang Y, Liang Q, Cui L, Wang L. The application of machine learning on antibody discovery and optimization. *Molecules* 2024;**29**:5923.
211. Hsieh K, Wang Y, Chen L, Zhao Z, Savitz S, Jiang X, et al. Drug repurposing for COVID-19 using graph neural network with genetic, mechanistic, and epidemiological validation. *Res Sq* 2020. rs.3.rs-114758.
212. Wang W, Feng S, Ye Z, Gao H, Lin J, Ouyang D. Prediction of lipid nanoparticles for mRNA vaccines by the machine learning algorithm. *Acta Pharm Sin B* 2022;**12**:2950−62.
213. Wang W, Chen K, Jiang T, Wu Y, Wu Z, Ying H, et al. Artificial intelligence-driven rational design of ionizable lipids for mRNA delivery. *Nat Commun* 2024;**15**:10804.
214. Ding DY, Zhang Y, Jia Y, Sun J. Machine learning-guided lipid nanoparticle design for mRNA delivery. *arXiv preprint* 2023. 2308.01402.
215. Moayedpour S, Broadbent J, Riahi S, Bailey M, V, Thu H, et al. Representations of lipid nanoparticles using large language models for transfection efficiency prediction. *Bioinformatics* 2024;**40**. btae342.
216. Witten J, Raji I, Manan RS, Beyer E, Bartlett S, Tang Y, et al. Artificial intelligence-guided design of lipid nanoparticles for pulmonary gene therapy. *Nat Biotechnol* 2024; **43**:1790−9.
217. Wu K, Yang X, Wang Z, Li N, Zhang J, Liu L. Data-balanced transformer for accelerated ionizable lipid nanoparticles screening in mRNA delivery. *Briefings Bioinf* 2024;**25**:bbae186.
218. Chan A, Kirtane AR, Qu QR, Huang X, Woo J, Subramanian DA, et al. Designing lipid nanoparticles using a transformer-based neural network. *Nat Nanotechnol* 2025;**20**:1491−501.
219. Ou Y, Zhao J, Tripp A, Rasoulianboroujeni M, Hernández-Lobato JM. A deep generative model for the design of synthesizable ionizable lipids. *arXiv preprint* 2024. 2412.00928.
220. Zhao J, Ou Y, Tripp A, Rasoulianboroujeni M, Hernández-Lobato JM. Generative model for synthesizing ionizable lipids: a monte carlo tree search approach. *arXiv preprint* 2024. 2412.00807.
221. Ogden PJ, Kelsic ED, Sinai S, Church GM. Comprehensive AAV capsid fitness landscape reveals a viral gene and enables machine-guided design. *Science* 2019;**366**:1139−43.
222. Bryant DH, Bashir A, Sinai S, Jain NK, Ogden PJ, Riley PF, et al. Deep diversification of an AAV capsid protein by machine learning. *Nat Biotechnol* 2021;**39**:691−6.
223. Zhu D, Brookes DH, Busia A, Carneiro A, Fannjiang C, Popova G, et al. Optimal trade-off control in machine learning-based library design, with application to adeno-associated virus (AAV) for gene therapy. *Sci Adv* 2024;**10**:eadj3786.
224. Eid F-E, Chen AT, Chan KY, Huang Q, Zheng Q, Tobey IG, et al. Systematic multi-trait AAV capsid engineering for efficient gene delivery. *Nat Commun* 2024;**15**:6602.
225. Guo B, Zheng H, Lin X, Jiang H, Mo A, Wei H, et al. Mapping AAV capsid sequences to functions through function-guided *in silico* evolution. *bioRxiv* 2024;**2024**:10.11.617764.
226. Guo Y, Shen Z, Zhao W, Lu J, Song Y, Shen L, et al. Rational identification of novel antibody-drug conjugate with high bystander killing effect against heterogeneous tumors. *Adv Sci* 2024;**11**:2306309.
227. Xu S, Xie L, Dai R, Lyu Z. Dumpling GNN: hybrid GNN enables better ADC payload activity prediction based on the chemical structure. *Int J Mol Sci* 2025;**26**:4859.
228. Angiolini L, Manetti F, Spiga O, Tafi A, Visibelli A, Petricci E. Machine learning for predicting the drug-to-antibody ratio (DAR) in the synthesis of antibody−drug conjugates (ADCs). *J Chem Inf Model* 2025;**65**:5847−55.
229. Su A, Luo Y, Zhang C, Duan H. Linker-GPT: design of Antibody−drug conjugates linkers with molecular generators and reinforcement learning. *Sci Rep* 2025;**15**:20525.
230. Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, et al. AlphaFold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res* 2024;**52**:D368−75.