

Structure prediction and engineering of *de novo* proteins

Lin Qi^{a,b}, Mark Isalan^{a,b,*}

^a Department of Life Sciences, Imperial College London, Exhibition Road, SW7 2AZ London, UK

^b Imperial College Centre for Engineering Biology, Imperial College London, SW7 2AZ, UK

ARTICLE INFO

Keywords:

Structure prediction
Protein engineering
Deep learning
de novo protein
Random sequences

ABSTRACT

De novo proteins represent a challenging frontier at the intersection of evolutionary biology and synthetic biology. Broadly, the term encompasses two distinct categories: naturally evolved *de novo* proteins, which arise from previously non-coding DNA, and synthetically designed *de novo* proteins, which are created from scratch through computational and experimental methods. This review provides an overview of the definitions, history, functional significance, and investigation methods for both categories. By examining potential challenges in current *de novo* protein research, this review highlights the growing convergence between natural evolutionary processes and rational protein engineering, reflecting their importance in biological discovery and human innovations.

1. Introduction

Proteins are fundamental to life, functioning as enzymes, structural elements, and regulators of numerous biological processes [1–5]. Recently, scientists have made significant efforts to uncover their folding mechanisms and evolutionary trajectories from first principles [6–10]. Traditionally, protein evolution was thought to rely on the duplication and divergence of pre-existing genes [11,12]. However, recent discoveries have shown that new protein-coding genes can emerge from previously non-coding regions [13,14]. These genes, which evolve from non-coding parts of the genome, are known as *de novo* genes. Under specific selective pressures, these ‘young’ genes can acquire functional roles, including helping bypass stop codons in *Saccharomyces cerevisiae* [15,16] and being essential for spermatogenesis in *Drosophila* [15,16].

In addition to studying *de novo* genes by analysing the transcriptomes and genomes of different organisms [17,18], scientists have also attempted to mimic *de novo* gene birth by constructing libraries of random sequences. Several studies have demonstrated that folded proteins can indeed be produced from random sequences [19,20], including the first *de novo* selection of a functional protein that bound ATP [19,20]. However, challenges remain in estimating the proportion of functional proteins within these random sequences and in validating the functions of *de novo* proteins.

In parallel, advances in structure prediction have enabled the rational and computational design of entirely new protein structures and

functions. The wealth of crystal structures and well-characterized physicochemical properties of proteins available in databases provide valuable training sets for understanding protein folding [21–23]. Protein engineering based on structural topologies learned from these large datasets is referred to as *de novo* protein design. In recent years, state-of-the-art structure prediction algorithms have achieved atomic-level accuracy [24–26], making *de novo* designed proteins closely resemble experimentally validated structures. From novel topologies to new functions, these artificially designed proteins have been successfully applied in chemical catalysis [27–29], therapeutic development [30], and beyond. Together with the naturally evolved *de novo* genes, these two developments currently frame the landscape of *de novo* proteins.

Despite the remarkable progress in *de novo* protein design, a major bottleneck remains: the relatively low success rate in designing functional *de novo* proteins. Functional designs often involve biomolecular interactions, particularly protein-ligand or protein-protein interactions, which introduce additional complexity and computational cost. This complexity reduces the accuracy of interaction predictions, thereby limiting the functional reliability of the designed proteins. Moreover, the exploration of *de novo* gene birth, by mimicking natural processes through random sequence libraries, remains limited because of the practical challenges of making very large libraries [31]. The fraction of functional proteins that can arise from such random sequences is still unclear, leaving a significant gap in our understanding. In this review, recent advances in *de novo* protein engineering are discussed from two complementary perspectives. The first focuses on artificially designed *de*

* Corresponding author at: Department of Life Sciences, Imperial College London, Exhibition Road, SW7 2AZ London, UK.

E-mail address: m.isalan@imperial.ac.uk (M. Isalan).

<https://doi.org/10.1016/j.bbapap.2025.141100>

Received 2 July 2025; Received in revised form 29 August 2025; Accepted 14 September 2025

Available online 15 September 2025

1570-9639/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

novo proteins, beginning with the history of structure prediction algorithms and examining how their development has influenced the success of *de novo* design. The second centers on evolved *de novo* proteins, explored through wet lab studies. This includes identifying natural *de novo* genes across diverse organisms using high-throughput sequencing, as well as investigating their emergence *via* random sequence libraries. Finally, we discuss the current limitations in the field and propose potential strategies to overcome these challenges in future research.

2. *De novo* protein design by structure prediction

Protein function is determined by structure, which is primarily dictated by amino acid sequence. As early as the 1970s, the question was raised as to whether protein structure can theoretically be predicted solely from amino acid sequence [32]. In response, researchers began to propose various algorithms to tackle this challenge [33–35]. The accuracy of these algorithms directly influences their utility in guiding wet lab experiments and, with the growing number of experimentally determined structures and the rapid advancement of computer science, structure prediction accuracy has improved significantly. Today, state-of-the-art prediction algorithms enable researchers not only to better understand protein folding but also to design proteins with entirely new folds and novel functions. These are explored in more detail in the following section.

2.1. History of structure prediction and early discovery for *de novo* protein design

The growing understanding of protein secondary structure in the 1950s–1970s laid the foundation for exploring protein 3D structures and developing structure prediction algorithms [36,37]. The earliest protein structure prediction method, homology modeling, was introduced in the 1980s [38] (Fig. 1). This approach is based on the principle that proteins with similar sequences tend to adopt similar structures. Homology modeling typically involves three major steps [39]: 1) The algorithm searches the Protein Data Bank (PDB) for known protein structures (templates) with similar sequences, using tools such as BLAST or HMM-based methods [40,41]. 2) Once suitable templates are identified, accurate sequence alignment is performed, using tools like Clustal Omega [42], to map residues. 3) A structural model is then constructed and refined through energy minimization. However, due to its dependence

on sequence similarity, homology modeling has limited applicability in *de novo* protein design. Its inability to predict entirely new folds, because of heavy reliance on existing structures in the PDB, makes it unsuitable for modeling proteins with low sequence identity (e.g., <30 %) [43]. Furthermore, *de novo* protein design often requires atomic-level accuracy, particularly in side-chain placement, which traditional homology modeling tools such as SWISS-MODEL cannot reliably achieve [39]. Although homology modeling is less commonly used today, it has been successfully applied in specific cases. For instance in engineering G protein-coupled receptors (GPCRs), to predict potential active sites and transmembrane topologies [44]. When combined with molecular docking, this method can also support virtual screening before experimental validation, helping to guide wet lab efforts [44].

To overcome the limitations of homology modeling, threading-based methods were introduced in the 1990s. These approaches are particularly suitable for predicting the structures of proteins with low sequence identity, as the diversity of known protein topologies is relatively limited (~1400 folds based on secondary structure content) [45], and proteins with remote homology can still adopt similar folds [46,47]. Threading-based structure prediction typically involves three main steps [48]: 1) A scoring function is developed to evaluate the compatibility between a given sequence and a library of structural templates. 2) The algorithm identifies the best sequence-structure alignments for each template and selects the top-ranked candidate. 3) A structural model is constructed and refined through energy minimization. Threading methods generally outperform homology modeling, especially in cases involving remote homologs [34,35]. A notable example of threading algorithm is Phyre2 [48], which was ranked among the top performing algorithms in the CASP9 competition [49]. Although threading methods are limited to predicting structures based on existing folds, they can still support protein design using selected scaffolds. For example, EvoDesign has been employed to engineer SARS-CoV-2 spike proteins with potentially enhanced immunogenicity [50,51]. However, threading-based methods are less effective for discovering proteins with entirely novel folds or functions, limiting their applicability in *de novo* protein design. Consequently, their use in real-world functional protein discovery remains constrained.

In the late 1990s, *ab initio* (or *de novo*) methods were introduced to predict protein structures, particularly for proteins with low or no detectable sequence homology [9,57]. These algorithms generally involve four major steps: 1) Design of an energy function, based on

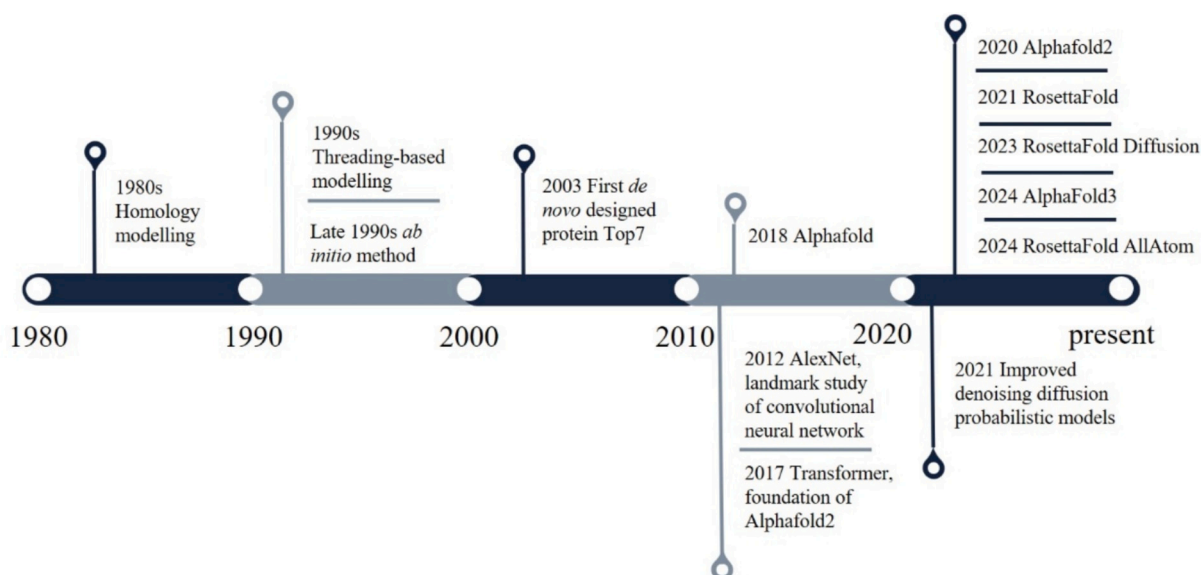


Fig. 1. History of structure prediction algorithms and development of *de novo* protein design. The development of protein structure prediction algorithms [24–26,39,48,52,57,63–65] is closely related to the advance of computational science [66–68].

physical principles or knowledge-based potentials, to identify the lowest energy state of a protein conformation. 2) Generation of short peptide fragments, typically extracted from known protein structures. 3) Assembly of full-length conformations using fragment insertion, guided by the designed energy function. 4) Optimization of geometry and side-chain placements for the final candidate structures. These breakthrough methods significantly improved the prediction accuracy for proteins with remote homology. A representative example is Rosetta [57], developed by David Baker's lab, which consistently ranked among the top-performing algorithms in the CASP (Critical Assessment of Structure Prediction) competitions [58] and outperformed all others for many years. Powered by *ab initio* approaches, *de novo* protein design achieved remarkable successes. In a landmark study, the novel protein Top7 was computationally designed and experimentally validated [52]. This 93-residue α/β protein adopted a completely new topology, comprising two α -helices packed against a five-stranded antiparallel β -sheet—a fold not found in nature. Its design model and experimentally determined X-ray structure matched with high precision (1.2 Å RMSD), demonstrating atomic-level accuracy in a *de novo* fold [52]. Furthermore, Top7 showed great thermal stability; it remained folded and structurally intact even at 98 °C. The successful design of Top7 further inspired creation of *de novo* proteins with novel folds, including various α -helical and β -strand architectures [59–61] (Table 1). For example, *de novo* α -helical bundles with diverse topologies exhibited exceptional thermodynamic stability, and experimental structural validation demonstrated that their crystal structures closely match the computational design models, particularly in core packing [53]. *De novo* designed β -barrel proteins demonstrated their activity in membrane insertion and fluorescence activation [55,62]. Even the more challenging α/β TIM-barrel fold has been designed with atomic-level accuracy [56]. More importantly, functional enzymes with no natural counterparts, such as Kemp eliminases [28] and retro-aldolases [29], were successfully engineered, demonstrating the power of structure-based design. For over a decade, *ab initio* methods were the most powerful tools for protein structure prediction. However, their high computational cost and limitations with larger proteins eventually led to their replacement by deep learning-based approaches, which now define a new era in *de novo* protein design. (See Table 2.)

2.2. Structure prediction algorithms with deep learning: A new era in *de novo* protein design

Since convolutional neural networks (CNNs) demonstrated remarkable success in image processing during the 2010s [68–70], the power of deep learning became widely recognized. This breakthrough led researchers to apply deep learning to various fields within the biological sciences, including medical diagnosis [71], bioimage analysis [70,72], and protein structure prediction [24–26,73] (Fig. 1). Progress in these applications has closely followed the development of cutting-edge algorithms. One of the earliest and most impactful deep learning-based structure prediction tools is AlphaFold [73], which was developed by Google DeepMind. AlphaFold made its debut in the CASP13 competition [74], where it achieved top ranking. It significantly outperformed traditional methods, particularly in cases involving proteins with no available structural templates. In the “no-template” category at CASP13,

Table 2
Summary of protein structure prediction algorithms.

Method	Type	Key features	References
Homology Modeling	Template-based method (TBM)	Uses known structures as templates, accuracy high when sequences have high sequence identity (>70 %) but drops when sequence identity below 30 %	[38,39]
Threading (Fold Recognition)	Template-based method (TBM)	Relies on fold database, predict distant homologs better than homology modeling, but unable to predict novel folds	[35,47]
<i>ab initio</i> Methods	Template-free method (FM)	Predicts structure from physical principles, able to predict novel folds but computationally expensive	[57,88]
AlphaFold2	Transformer-based method	Uses MSAs and EvoFormer; reaches near-experimental accuracy but limited by disordered regions and proteins with low database homology	[24]
ESMFold	Large language model-based	Uses single-sequence Transformer, enables fast and high throughput structure prediction	[89]
D-I-TASSER	Deep learning-based method	Uses deep learning contact/distance/hydrogen-bond restraints and domain handling	[82]
Boltz-2	Transformer-based method	Predicts biomolecular complex structures and binding affinities; achieves near free-energy perturbation (FEP) accuracy in affinity predictions while being ~1000× faster	[98]

AlphaFold achieved the highest accuracy for 25 out of 43 targets, while the second-best method succeeded in only 3, clearly highlighting the superiority of AlphaFold [63]. Although the original version of AlphaFold was not directly applied to *de novo* protein design, its success validated the potential of neural networks in biological research and sparked a wave of interest and innovation in applying deep learning to structural biology and related fields.

In 2017, Google DeepMind introduced the Transformer architecture in the landmark paper “Attention Is All You Need” [67] (Fig. 1). This novel framework fundamentally reshaped deep learning by replacing traditional recurrence and convolutional mechanisms with self-attention [67]. Compared to conventional neural networks [68], self-attention enables the model to directly capture long-range dependencies while allowing for efficient parallelization, significantly increasing computational speed on GPUs [67]. Building on this innovation, DeepMind developed AlphaFold2 [24], a transformative protein structure prediction algorithm that redefined computational biology.

Table 1
Examples of *de novo* designed proteins.

Protein name/type	Algorithm	RMSD to crystal structure (C α)	Initial design	Experimentally tested design	References
Top7	Rosetta fragment assembly	~1.2 Å	172	1 structural validated	[52]
Three-helix bundles	Crick coiled-coil equations and Rosetta	~1.1 Å	N/A	4 tested, 3 structural validated	[53]
Four-helix bundles	Crick coiled-coil equations and Rosetta	~0.7 Å	N/A	3 tested, 1 structural validated	[53]
Five-helix bundles	Crick coiled-coil equations and Rosetta	~0.4 Å	N/A	1 structural validated	[53]
Self-assembling helical filaments	Rosetta design of oligomeric interfaces	~0.9–3.4 Å	23,000	124 tested, 6 structural validated	[54]
Transmembrane β -barrel	Rosetta fragment assembly	~1.1 Å	200	66 tested, 2 structural validated	[55]
Designed TIM-barrel	Rosetta symmetric-based design	~1.28 Å	N/A	33 tested, 1 structural validated	[56]

The AlphaFold2 architecture consists of three major components [24]: 1) An Input module, which processes multiple sequence alignments (MSAs) and initial pairwise features. 2) Evoformer, which is a transformer-inspired core network composed of 48 iterative blocks that refine both the MSA and pair representations. 3) A Structure module, which transforms the processed features into a full 3D atomic representation of the protein. This model also includes a recycling mechanism, in which the output is reintroduced into the Evoformer and structure module for further refinement. AlphaFold2 dominated CASP14, achieving top accuracy for 88 out of 97 targets, and reaching atomic-level precision for most [75,76]. The release of AlphaFold2 through Google Colab further democratized access to accurate structure prediction, enabling researchers worldwide to model protein structures efficiently and affordably. As a result, deep learning has become central

to both protein structure prediction and *de novo* protein design, now increasingly integrated with experimental workflows to accelerate the discovery process. For example, David Baker and colleagues have implemented AlphaFold2 to design DNA-binding proteins with customized binding sites and affinities [77]. The high-throughput and efficient structure prediction by AlphaFold2 has significantly reduced the cost of wet lab experiments.

Inspired by the transformer-based architecture, Baker's group also introduced RoseTTAFold in 2021 [64]. This upgraded structure prediction framework incorporates a three-track network [64]: 1) A 1D track that processes sequence-level information and MSA profiles. 2) A 2D track that captures residue-residue pairwise features, and 3) a 3D track that manages spatial coordinate predictions of backbone atoms. RoseTTAFold nearly matched AlphaFold2's performance in CASP14 and

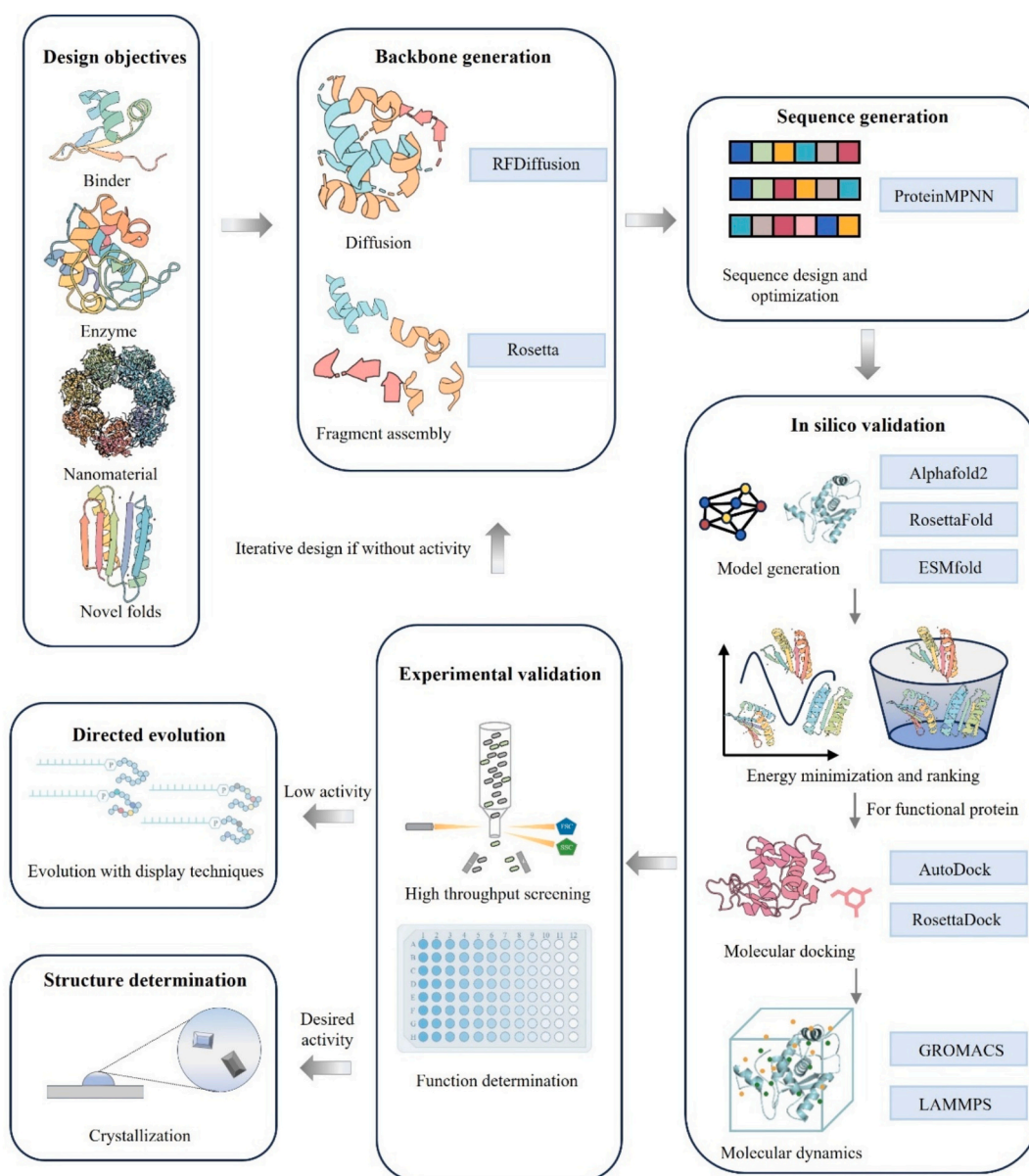


Fig. 2. Schematic Figure of *de novo* protein design workflow. The *de novo* protein design process begins with defining the design objective such as creating a binder for other proteins or ligands, an enzyme, or other nanomaterial. Backbone structures are then generated using Rosetta [88] or RFDiffusion [65], followed by amino acid sequence design using ProteinMPNN [94]. Designed sequences undergo structural prediction [24,64,89] and functional assessment via molecular docking [112,113] and molecular dynamics simulations [24,114]. Candidates are experimentally validated through methods such as yeast display screening and Surface plasmon resonance (SPR) binding affinity characterization. Finally, candidates without activity require iterative design and candidates with low activity require directed evolution [28]. Successful candidates are experimentally validated through protein purification, structure determination and other biochemical characterization.

has been successfully used to design diverse protein topologies, including β -barrel proteins [78], homodimers [79], and heterotrimers [80]. More importantly, it has enabled the design of functional proteins, such as high-affinity protein binders [81]. Recently, Yang Zhang's group proposed D-I-TASSER, which integrates deep learning-derived spatial restraints with iterative threading-based fragment assembly simulations [82]. Specifically, the prediction begins with 1) a deep multiple sequence alignment *via* DeepMSA2 [83]; 2) then the spatial restraints including inter-residue contacts, distance maps, and hydrogen-bonding networks are predicted using DeepPotential and AttentionPotential [84]; 3) after identification of structural templates using LOMETS3 [85], the final models are generated by Replica-Exchange Monte Carlo (REMC) fragment-assembly simulations [86]. In CASP15, D-I-TASSER showed outstanding performance in large and multidomain proteins [87]. These algorithms further demonstrate the potential of attention mechanisms in structure prediction as well as their power in *de novo* mini-protein design.

Despite the revolutionary advancements brought about by algorithms such as AlphaFold2 [24], functional *de novo* protein design remains limited by a low success rate. Functional proteins typically participate in complex biomolecular interactions, such as protein–protein or protein–ligand binding. Therefore, effective design pipelines must integrate multiple steps including structure prediction, molecular docking, and binding affinity scoring (Fig. 2). Errors in any step can lead to functional failure. Additionally, proteins often undergo conformational changes upon binding, and the dynamic nature of both proteins and ligands significantly increases the complexity of accurate predictions. Recognizing these challenges, both DeepMind and David Baker's lab began exploring the prediction of protein–ligand complexes after CASP13.

In 2020, diffusion models emerged as a promising solution [66] (Fig. 1). These probabilistic generative frameworks learn to denoise data through iterative refinement and were originally developed for applications in image processing [66,90]. Their success in that domain inspired their applications in structural biology. In protein modeling, diffusion models can operate on representations such as torsion angles and coordinate graphs, or atomic point clouds, generating accurate 3D structures *via* a learned reverse-noising process [91–93]. By simulating a transition from disorder to order, these models mirror the process of protein folding, refining coarse structural predictions into precise conformations. Building on this innovation, DeepMind introduced AlphaFold3, which incorporates a diffusion model into its structure refinement module [25]. Another key architectural change in AlphaFold3 is the replacement of the Evoformer with the Pairformer, enabling greater efficiency and scalability by using single-sequence embeddings instead of computationally expensive MSA-based attention [25]. In addition to achieving protein structure prediction accuracy comparable to AlphaFold2, AlphaFold3 offers enhanced capabilities, including the modeling of post-translational modifications (e.g., phosphorylation, glycosylation) and interactions with DNA, RNA, small molecules, and ions [25]. The accurate prediction of protein–ligand interactions and multimeric complexes (binding proteins or 'binders') provided by AlphaFold3 opens new avenues for *de novo* protein design, bringing the field closer to reliably engineering proteins with novel and specific functions. Despite there being no current examples of AlphaFold3 designed *de novo* proteins (due to its recent release), a full demonstration of a *de novo* designed binder is anticipated soon.

Similarly, the diffusion model was integrated into the RosettaFold framework, resulting in the development of RosettaFold Diffusion in 2023 [65]. This generative model, based on denoising diffusion probabilistic models (DDPMs) [66], enables the iterative generation of complete protein backbones. RosettaFold Diffusion significantly improved the success rate of *de novo* binder design, raising it to approximately 19 %, a notable advancement over previous approaches [65,92]. This improvement was in part achieved by coupling the diffusion-based backbone generation with advanced sequence design algorithms such

as ProteinMPNN [94]. ProteinMPNN is a deep learning-based message-passing neural network, introduced in 2022, that optimizes amino acid sequences for a given protein structure, thereby dramatically increasing the likelihood that a designed backbone will fold into a stable, functional protein. By integrating such sequence refinement steps, the RosettaFold Diffusion pipeline can more reliably produce effective protein designs [65], contributing to the higher experimental success rates observed in practice. To further enhance the prediction of protein–ligand interactions and optimize *de novo* protein design, RosettaFold All-Atom (RFAA) was introduced in 2024 [26]. Unlike earlier approaches that predicted backbone and side-chain torsion angles for subsequent structure reconstruction, RFAA directly generates full-atom coordinates, including side chains and torsional angles, with high precision [26]. More importantly, RFAA extends its predictive capabilities to include molecules beyond proteins, such as nucleic acids, metal ions, and post-translationally modified residues, enabling atomic-level modeling of complex biomolecular systems [26]. Powered by the combined strengths of RosettaFold Diffusion and RFAA, the field has witnessed a notable surge in the successful design of functional *de novo* proteins. For example, RosettaFold Diffusion has been applied to engineer proteins capable of neutralizing snake venom, achieving a binding affinity of approximately 0.9 nM [30]. In another case, a *de novo* designed enzyme generated using RosettaFold Diffusion was shown to catalyze the hydrolysis of serine [27], demonstrating its potential in biocatalysis. These breakthroughs in functional designs show that *de novo* proteins can now deliver lifesaving and industrially relevant proteins, proving that *de novo* design is shifting towards real-world activities (Table 3). Building on these structural advances, computational function-prediction algorithms developed recently also play a critical role in validating and guiding design outcomes. Established tools like CLEAN use predicted structures to infer function—providing annotations for ligand-binding sites, Gene Ontology (GO) terms, and Enzyme Commission (EC) numbers based on structural similarity to known proteins [95]. For detecting potential functional sites even in novel folds, ScanNet compares structural motifs irrespective of sequence homology to predict binding pockets and ligand associations [96]. Together, these function-prediction frameworks—paired with precise backbone generation *via* diffusion models and all-atom coordinate modeling—form an integrated pipeline enabling design, assessment, and functional realization of novel *de novo* proteins.

In addition to the structure prediction algorithms developed by Google DeepMind and the Baker Lab, Jack and colleagues have recently introduced Boltz-1, which employs an AlphaFold 3-inspired co-folding transformer architecture [97]. As an open-source structure prediction algorithm, Boltz-1 achieved AlphaFold3-level performance for complex biomolecular structure prediction [97]. Building on this, Boltz-2 further added an affinity prediction module, enabling the joint modeling of structure and binding affinity [98]. It approached the accuracy of physics-based free energy perturbation (FEP) methods but at over 1000 $\times$ of the speed, and outperformed all existing methods in machine learning and docking methods [98]. With improved controllability and integration of molecular dynamics data, high throughput *in silico* ligand screening could be achieved by Boltz-2. Apart from that, Meta AI introduced a novel language model-based approach called ESMFold in 2023 [89]. Inspired by the rapid progress in natural language processing (NLP) [99,100], researchers have begun applying language models to the biological sciences, giving rise to protein language models [89,101]. Meta AI's ESM-2 is a transformer-based protein language model analogous to BERT [102], employing self-attention mechanisms with rotary positional embeddings (RoPE) [103]. The model was trained using UniRef50 [104,105], which is a clustered, non-redundant collection of UniProt sequences [23]. ESM-2 supports high-accuracy predictions of protein functions, including Gene Ontology (GO) annotations [106]. Building on ESM-2, the derivative algorithm ESMFold translates learned sequence embeddings directly into 3D structural coordinates, achieving prediction accuracy comparable to AlphaFold2. Notably, ESMFold does

Table 3Examples of *de novo* designed functional proteins.

Protein name/type	Algorithm	Experimentally tested designs	Designs with activity	Outcome	References
Kemp eliminases	Rosetta active-site design	59 designs	8 designs	$6\text{--}160\text{M}^{-1}\text{s}^{-1}$	[28]
High-affinity binders of helical peptides	Rosetta	PTH&GCG&NPY: 12 designs	PTH: 3 designs; GCG: 4 designs; NPY: 8 designs	Picomolar to sub nanomolar binding affinities	[111]
	RF _{joint}	PTH: 192 designs; GCG&NPY: 96 designs	PTH: 44 designs; GCG&NPY: N/A		
	Hallucination	46 designs	4 designs		
Snake venom binder	RF Diffusion	PTH&Bim: 96 designs	PTH: 56 designs; Bim: 25 designs	80 %–100 % <i>in vivo</i> protection rate	[30]
	RF Diffusion	ScNtx binder: 44 designs, α -cobratoxin: 42 designs, cytotoxin: 55 designs	For each target, 1 showed initial activity		
Serine hydrolase	RF Diffusion	After two rounds of filtering, 132 proteins were designs in round 3	20 designs showed 4MU substrates hydrolysis activity	Most efficient design showed a $k_{\text{cat}}/K_{\text{m}}$ of $2.2 \times 10^5\text{M}^{-1}\text{s}^{-1}$	[27]

not rely on multiple sequence alignments (MSAs), making it particularly advantageous for modeling *de novo* proteins, artificial sequences, and understudied proteins from metagenomic sources [89]. By eliminating the alignment step, ESMFold also enables fast, high-throughput structure prediction at a genomic scale. Beyond prediction, ESMFold is now being used for inverse protein design, as well as protein hallucination [107,108], where structural targets are used to generate compatible *de novo* sequences. Additionally, Meta AI has developed prompt-based design pipelines, using ESM-2 in conjunction with ESMFold to generate structurally viable *de novo* proteins from high-level functional or structural prompts [109,110], opening new possibilities for automated and user-guided protein design.

2.3. Conclusion: *de novo* protein design

De novo protein design stands at the forefront of synthetic biology and protein engineering because it enables the creation of entirely novel proteins that are not derivable from existing natural sequences. This capability is more than a remarkable intellectual feat; it transforms our understanding of protein folding and function into an active engineering discipline. Rather than merely studying what nature has already evolved, researchers can now articulate desired structural or functional outcomes and deliberately construct proteins that fulfill those specifications.

A typical *de novo* protein design workflow reflects this goal-oriented engineering approach (Fig. 2). It begins with clear design objectives, such as a desired fold or specific functional behavior. Researchers then use generative models, including diffusion models [65], to create candidate backbones. Sequences are assigned using inverse folding methods, such as ProteinMPNN, which predicts amino acid sequences that are thermodynamically plausible and likely to fold into the specified backbone [94]. The structural integrity of these constructs is evaluated with prediction tools like AlphaFold2, RoseTTAFold, or ESMFold [24,64,89]. When the target involves binding or function, molecular docking and molecular dynamics simulations are often performed to assess interaction fidelity *in silico* [112–114]. Promising candidates are then expressed and examined experimentally through structural determination methods (e.g., X-ray crystallography, CryoEM) and functional assays (e.g., SPR) [30,111]. If they fall short of expectations, iterative refinement *via* redesign or directed evolution (e.g., error-prone PCR) is pursued [28].

De novo protein design is a fast-evolving field, already with significant achievements in creating novel therapeutic proteins and industrial enzymes. The development of this field is also closely associated with the advances of computer science, particularly in the areas of deep learning. However, despite these achievements, several critical challenges still constrain the broader success and applicability of *de novo* protein design. First, with limited understanding of

sequence–structure–function relationships, the success rate of designing functional proteins is still quite low, and the designed proteins may lack desired activities or suffer from problems such as low expression levels and low solubility. Although solubility prediction algorithms such as CamSol, DeepSol, and NetSolP have been proposed to help filter and identify more promising candidates [115–117], the activity of *de novo* designed proteins generally remains low in most cases. Designs that lack desired activity generally require directed evolution or iterative designs to further improve the activity. For example, in the recent *de novo* design of snake venom binder, only 3 designs demonstrated initial activity, and iterative design was needed for increasing the binding affinity [27]. Another study reported ~2 % success rate of $\alpha\beta\beta\alpha$ fold protein and proposed that fold complexity greatly influences design success, with highly challenging targets resulting in very low hit rates [118]. Second, although structure prediction algorithms such as AlphaFold2 and RoseTTAFold have revolutionized the field, they remain constrained by their training data [24,64]. These models struggle with novel folds, unconventional sequences, or proteins with previously unseen functionalities. In addition, although the Protein Data Bank (PDB) contains a large number of high-resolution crystal structures [21,22], the number of protein–protein and protein–ligand complexes is still limited, leading to the insufficient prediction accuracy of biomolecular interactions. Third, negative feedback from failures in expression, solubility, or folding is often poorly analyzed and not reincorporated into models, limiting the opportunity to improve models and design strategies based on unsuccessful outcomes.

3. Naturally evolved *de novo* proteins versus *de novo* proteins from random sequence libraries

Over the past two decades, computational protein design has explored the vast potential of sequence space and demonstrated that new-to-nature proteins can not only be structurally stable but also functionally active [27,30,52,79,107,108]. Apart from discovering novel folds or functions through such artificial design, researchers have also investigated *de novo* proteins experimentally, particularly those that arise naturally. Although non-coding regions of DNA were once dismissed as “junk DNA” [11], they are now recognized as potential sources of novel biological functions [17]. Through a combination of mutations, transcriptional activation, and the acquisition of translation initiation signals, non-coding sequences can evolve into functional genes [119]. These naturally evolved *de novo* genes provide valuable insights into gene emergence mechanisms and the evolution of protein structure and function. Researchers typically explore naturally occurring *de novo* genes through two main approaches: 1) identification through high-throughput sequencing and bioinformatics analysis, followed by expression assays and functional characterization to validate their biological relevance [17,120]. 2) Experimental simulation of *de novo* gene

birth by constructing random sequence libraries, allowing the characterization of novel, functional proteins that may mimic natural evolutionary pathways [19,20]. In the sections below, we will examine both of these in turn.

3.1. Naturally evolved *de novo* proteins

In the early 1970s, the possibility of *de novo* gene birth sparked considerable debate [11,12,121]. François Jacob famously argued that the emergence of novel genes from random associations of amino acids was highly unlikely [11]. However, other scientists held opposing views, and early experimental evidence suggested that novel proteins could indeed arise from previously non-coding regions of DNA [19,121]. In the mid-2000s, transcriptomic studies and bioinformatics analyses identified some of the first putative *de novo* genes in *Drosophila yakuba* and *Drosophila erecta* [13,14]. Although functional validation remained challenging, these studies proposed that *de novo* genes may only remain detectable for a short evolutionary window, as they are prone to rapid divergence or loss [13]. Since then, a growing number of *de novo* genes have been discovered across various taxa [122,123], and the functions of several have been experimentally validated [15,16,124]. For instance, AGFP, a gene identified in cold-adapted fish, produces a protein that inhibits ice crystal growth, providing a vital survival advantage in freezing environments, yet it is absent in species not exposed to such pressures [124]. Functional *de novo* genes have also been reported in plants [125], mice [126], and humans [127], collectively supporting the view that *de novo* gene birth is a real and perhaps not uncommon evolutionary phenomenon.

To gain deeper insights into the mechanisms of *de novo* gene birth, researchers have also examined the structures of *de novo* proteins. One well-characterized example is BSC4 [122], a *de novo* gene from *Saccharomyces cerevisiae* involved in stop-codon suppression and DNA repair under nutrient-limited conditions [15,122]. Structural studies revealed that BSC4 adopts β -sheet structures and possesses a hydrophobic core, existing in a continuum of oligomeric states [15]. Based on these findings, researchers proposed that *de novo* proteins may initially adopt flexible, intermediate conformations before gradually evolving into more stable and well-defined folds [15]. Another structurally

studied *de novo* protein is Goddard, whose gene (*goddard*) is important for male fertility in *Drosophila* [123]. Combining computational modeling, molecular dynamics simulations, and experimental techniques such as nuclear magnetic resonance (NMR) and circular dichroism (CD), researchers found that Goddard contains a central α -helix (~20–30 residues) flanked by intrinsically disordered regions [16]. These findings suggest that only minor structural changes may occur after the initial emergence of certain *de novo* proteins [16]. This hypothesis was further supported by a recent large-scale analysis using advanced structure prediction tools such as AlphaFold2 and ESM-2. In a study of 555 *de novo* gene candidates in *Drosophila*, the majority of the encoded proteins were predicted to be poorly folded [128]. However, among the subset of well-folded candidates, researchers observed that major structural changes after gene birth were unlikely. Interestingly, other properties such as GC content, evolutionary rate, and expression patterns tended to shift over time, suggesting ongoing optimization through natural selection [128]. These structural studies collectively challenge the hypothesis that *de novo* proteins are merely random, non-functional peptides. Instead, they demonstrate that some *de novo* proteins can adopt functional and structured forms, potentially from the very moment they arise.

Nonetheless, the structural determination of *de novo* proteins remains both rare and technically challenging, even with machine learning tools (Fig. 3). Structural studies have often reported unsuccessful attempts at crystallizing *de novo* proteins, and several factors likely contribute to these difficulties. First, because *de novo* proteins typically originate from non-coding DNA, they often contain intrinsically disordered regions (IDRs) or are largely unstructured [16,128]. Such proteins may lack stable tertiary structures, making them poorly suited for high-resolution techniques like X-ray crystallography or nuclear magnetic resonance (NMR). Second, *de novo* proteins tend to be small, thermodynamically unstable, and prone to aggregation or misfolding. These properties often result in low expression yields and complicate *in vitro* purification, limiting the availability of sufficient material for structural analysis. Lastly, structural biology techniques like X-ray crystallography and NMR are time-consuming, expensive, and require relatively stable protein samples. As a result, research labs typically prioritize evolutionarily conserved, well-folded proteins with

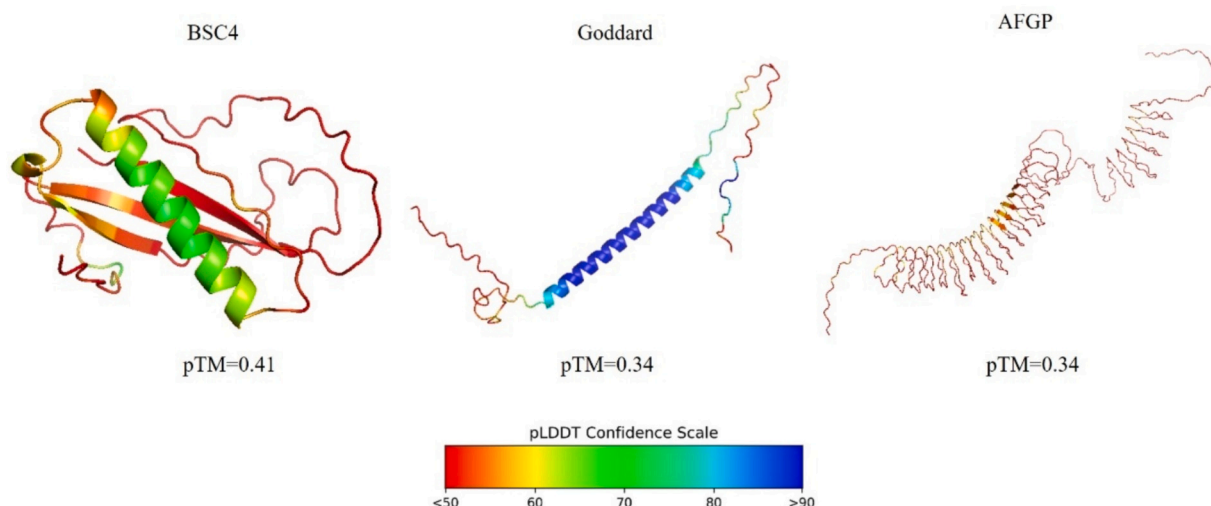


Fig. 3. Structures of naturally evolved *de novo* proteins predicted by AlphaFold2 [24]. The prediction confidence is presented by pLDDT (predicted Local Distance Difference Test) and pTM (predicted Template Modeling score) [129,130]. A pTM score higher than 0.8 indicates a high-confidence prediction, suggesting that the overall fold of the protein is likely correct. A pTM score between 0.5 and 0.8 represents moderate confidence, meaning that, while some structural elements may be accurate, the global topology is uncertain. A pTM score below 0.5 suggests a low-confidence prediction, where the overall structure is likely incorrect or unreliable [130]. A pLDDT higher than 90 indicates very high confidence, where both backbone and side-chain predictions are likely accurate. A pLDDT between 70 and 90 corresponds to high confidence, suggesting a reliable backbone structure with possible side-chain flexibility. A pLDDT between 50 and 70 reflects moderate confidence, where backbone predictions may contain significant errors. A pLDDT below 50 is considered low confidence, and the predicted structure is likely unreliable or incorrect [129]. These predicted *de novo* proteins therefore have low prediction confidence.

known functions and homologs, making *de novo* proteins a lower priority for structural characterization.

Due to the development of next-generation sequencing (NGS) technologies, naturally evolved *de novo* genes have been identified on a large scale, prompting researchers to investigate the frequency of *de novo* gene emergence. In a landmark study, Anne-Ruxandra Carvunis and colleagues employed RNA-seq and Ribo-seq to investigate the frequency of *de novo* gene emergence [17]. Their analysis identified approximately 1900 candidate proto-genes among *Saccharomyces cerevisiae* open reading frames (ORFs). Surprisingly, during the same evolutionary timeframe, more *de novo* ORFs were identified than novel genes resulting from gene duplication [17]. This finding suggested that *de novo* gene birth may hold evolutionary significance comparable to gene duplication, which was traditionally considered the dominant mechanism of novel gene formation. Another study applied deep RNA sequencing and ribosome profiling to examine *de novo* gene emergence in *S. cerevisiae*, under both rich medium and oxidative stress conditions [120]. The researchers identified 213 *de novo* transcripts, of which 97 were actively translated, indicating potential functional relevance [120]. Further analysis revealed that *de novo* proteins tend to be short in length and positively charged [120]. These high-throughput investigations supported the hypothesis that *de novo* gene birth is not a rare event, but a frequent and dynamic evolutionary process that can contribute significantly to the emergence of new functions.

3.2. Exploration of *de novo* proteins through random sequence libraries

To explore the intrinsic potential of raw sequence space to produce functional and foldable proteins, scientists have attempted to mimic *de novo* gene birth by constructing libraries of random protein sequences. One of the earliest such efforts was carried out by Davidson and colleagues, who hypothesized that, although naturally evolved *de novo* proteins are rare, folded proteins may be generated more frequently within random sequence libraries than previously assumed [19]. To test this hypothesis, they constructed a library of proteins composed exclusively of three amino acids: glutamine (Q), arginine (R), and lysine (L)—forming sequences 80 to 100 amino acids in length [19]. Remarkably, they found that approximately 5 % of the QRL library could be successfully expressed in *E. coli*, providing early evidence that even simplified random sequences can yield soluble, expressible proteins [19]. Further support came from another study by Chiarabelli and colleagues, who created a 50-residue random sequence library centered around a thrombin recognition site [131,132]. Using phage display selection, they discovered that roughly 20 % of the proteins in the library were capable of shielding the thrombin cleavage site, suggesting a degree of functional structure [131,132]. Circular dichroism (CD) spectroscopy confirmed that many of these selected proteins exhibited stable tertiary structures and were capable of reversible folding and unfolding [131,132]. These studies suggest that functional and foldable proteins may not be as rare in random sequence space as once believed, offering valuable insights into the plausibility of *de novo* protein emergence.

In the following years, researchers further investigated the biophysical properties of proteins generated from random sequence libraries, with a particular focus on solubility. For example, Prijambada and colleagues constructed a library of proteins consisting of 141 amino acids, of which 95 positions were fully randomized [133]. Their study found that approximately 20 % of the sequences produced soluble proteins, reinforcing the idea that a substantial fraction of random sequences can yield foldable, expressible products [133]. Building on this, the Yanagawa Group examined how amino acid composition affects solubility [134]. They constructed random sequence libraries comprising 5, 12, or 20 amino acids, respectively [134]. Their findings revealed that proteins generated from the 20-amino-acid library were more prone to aggregation than those from the more compositionally restricted libraries [134]. This suggested that limiting amino acid diversity may enhance solubility and reduce aggregation tendencies in

randomly generated proteins. Similarly, Newton and colleagues explored this concept in the context of evolutionary history, constructing random sequence libraries composed of 5, 9, 16, and 20 amino acids, respectively [135]. They also found that proteins expressed from the full 20-amino-acid library exhibited lower solubility compared to those from more constrained libraries. This reduced solubility was likely attributable to the inclusion of aromatic residues, which are known to promote hydrophobic interactions and aggregation [135]. These findings highlight how amino acid composition and evolutionary constraints can influence the biophysical properties of *de novo* proteins, indicating how nature may have selected for certain features during the early evolution of functional proteins.

The fraction of functional proteins within random sequence space has also been investigated by constructing random sequence libraries and experimentally enriching potential functional candidates through selection or screening methods. In a landmark study, Jack Szostak and colleagues explored *de novo* proteins using directed evolution techniques [20]. To mimic the process of *de novo* gene birth, they constructed a library of random 80-amino-acid length sequences. Through 18 rounds of selection using mRNA display, they successfully identified a family of ATP-binding proteins. These *de novo* proteins exhibited a binding affinity of approximately 100 nM for ATP, while showing minimal affinity for ATP analogs, indicating a degree of specificity despite their random origin [20]. The following studies further elucidate the structural features and evolutionary trajectories of this artificial ATP-binding protein [136–138]. They reveal that folding is stabilized both by burial of the hydrophobic core and by tetrahedral coordination of a zinc ion through a typical CXXCnCXXC zinc-binding motif [138]. 12 amino acid substitutions, introduced through *in vitro* evolution with mRNA display under denaturing-containing conditions, further enhance ATP-binding activity (e.g. R19W and K21E), solubility (e.g. S41Y), and thermal stability (e.g. N29T) [136]. Based on their results, the authors estimated that the frequency of functional ATP-binding proteins in random sequence space is approximately 1×10^{11} [20]. The direct evidence that function can emerge without evolutionary history demonstrated that functional sequences are not vanishingly rare and can be rapidly improved with the power of *in vitro* selection. More importantly, the non-biological scaffold can help to understand the origin of life and protein engineering from scratch. To date, this remains one of the only successful demonstrations of isolating functional proteins from a fully random sequence library using directed evolution. Despite the development of numerous high-throughput selection techniques, and the advent of next-generation sequencing, this approach has seen limited follow-up over the past two decades. This is likely due to the widespread assumption that the frequency of functional proteins in random sequence space is prohibitively low, making such efforts inefficient and resource intensive.

In addition to wet lab studies, the properties of random sequence libraries have also been investigated using computational approaches, although the conclusions drawn from these studies are sometimes contradictory. JiaFeng Yu and colleagues constructed two datasets of random protein sequences: the first random sequence library was generated based on the frequency of the occurrence of the twenty amino acids in the codon table, while the second assumed a uniform 5 % frequency for each amino acid. Their analysis showed that proteins generated using natural amino acid frequencies are generally more intrinsically disordered than those derived from uniformly distributed random sequences [139]. They attributed this to evolutionary selection favoring solubility and functional versatility, which may explain why natural proteins tend to exhibit greater dynamic properties than unevolved sequences. Another study by Annamária F. Ángyán and colleagues further explored the relationship between GC content and the disorder propensity of random sequences [140]. They found that higher GC content is generally associated with increased intrinsic disorder. However, in contrast to both experimental evidence and some computational predictions, they concluded that *de novo* proteins are not

necessarily more aggregation-prone than natural proteins [140]. More recently, the Hlouchová group employed a combined computational and experimental approach to compare random polypeptides with biological proteins. Using bioinformatic tools, they predicted the secondary structure content, disorder levels, and aggregation propensities of a library of 100-residue random sequences, followed by experimental characterization of several selected proteins [141]. Surprisingly, the secondary structure content and general physicochemical properties of random sequences were found to be broadly similar to those of natural proteins. Interestingly, the most soluble polypeptides were those with the least secondary structure and the highest disorder, likely due to their reduced tendency to aggregate [141]. While these studies provide useful insights into the relationship between sequence randomness, structural disorder, and solubility, their findings are largely based on computational predictions and lack full experimental validation across broader sequence spaces. As a result, it remains an open question as to what is the spectrum of features governing the behavior of random or *de novo* proteins, particularly for solubility.

3.3. Conclusion: natural *de novo* proteins versus random sequence libraries

Naturally evolved *de novo* proteins, which arise from previously non-coding DNA sequences, represent a fascinating class of proteins with profound implications for evolutionary biology, functional genomics, and synthetic biology. Although growing evidence supports their existence and functional relevance [15,16], our understanding of these proteins remains limited. Several challenges contribute to this knowledge gap. First, *de novo* proteins are typically small, thermodynamically unstable, and often contain intrinsically disordered regions, which makes them difficult to detect using conventional techniques [15,16]. Second, the scarcity of detailed characterization data hinders our ability to determine how these proteins emerge and evolve. Third, the lack of structural insights impedes a clear understanding of their biological mechanisms and functional integration within cellular systems.

One promising approach to better understand *de novo* proteins involves exploring random sequence space through generating and screening synthetic or randomized polypeptides to uncover novel functional properties [20,142]. This strategy allows researchers to investigate the limits of protein evolution and design, offering a window into the types of sequences that can give rise to foldable, functional proteins. However, this field is still in its infancy and faces significant challenges. Functional proteins such as protein-protein binders or enzymes are extremely rare in random sequence libraries and there remains a poor understanding of how biological function emerges from random, unevolved sequences. Moreover, we currently lack comprehensive maps that describe how key biophysical properties such as solubility, foldability, and stability are distributed across sequence space. Identifying and validating the rare functional hits requires labor-intensive downstream processes, including protein purification, activity assays, and structural determination. Furthermore, structure prediction tools, such as AlphaFold2 [24] and others, tend to perform poorly on unevolved or highly disordered sequences, limiting their utility in guiding experimental screening of *de novo* proteins.

4. Future directions

Two avenues of *de novo* protein discovery, *de novo* protein design and natural or directed evolution of such proteins, connect synthetic biology, molecular evolution, and origin-of-life studies. With the rapid advancement of computer science, AI for science is becoming increasingly influential. The application of artificial intelligence to *de novo* protein discovery is particularly promising, as it can reduce manual labor, streamline design pipelines, and enhance prediction accuracy. At the same time, high-throughput sequencing and screening in laboratory settings remain essential, as these can uncover generalizable rules

governing how new proteins emerge, evolve, and functionalize in biological processes.

The future of *de novo* protein investigation is likely to concentrate on a coordinated approach, combining both experimental and computational methods. For naturally evolved *de novo* proteins, it will be crucial to develop rigorous and standardized pipelines for gene identification, which can minimize false positives by ensuring that putative *de novo* proteins truly lack homologs in known databases. Furthermore, structural characterization of these proteins will be vital to illustrate the mechanisms underlying their biological activities. To mimic *de novo* gene birth, directed evolution techniques such as mRNA display and CIS display can offer powerful tools due to their capacity for screening ultra-diverse libraries with high efficiency [31,143–146]. These methods can identify rare functional proteins from random sequence space, providing insights into early protein evolution and novel functionalities. To enhance the efficiency and reproducibility of this research, incorporating classic synthetic biology frameworks, such as the Design-Build-Test-Learn (DBTL) cycle, is likely to be essential [147]. When integrated with laboratory automation [148–151], the DBTL approach can accelerate discovery and provide a systematic platform for engineering, screening, and refining *de novo* proteins.

Ultimately, the integration of high-throughput experimental screening with computational tools, including structure prediction, secondary structure estimation, and solubility prediction, may yield the most comprehensive and scalable approach to understanding and discovering *de novo* proteins. This multidisciplinary strategy will be critical for unlocking the full potential of *de novo* protein design and discovery, in both basic science and applied biotechnology.

CRedit authorship contribution statement

Lin Qi: Writing – original draft. **Mark Isalan:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

Mark Isalan reports financial support was provided by Office of Naval Research Global. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

MI is funded by an award from the Office of Naval Research Global (ONRG N62909-23-1-2099).

References

- [1] P.F. Johnson, S.L. McKnight, Eukaryotic transcriptional regulatory proteins, *Annu. Rev. Biochem.* 58 (1) (1989) 799–839.
- [2] S.A. Lambert, A. Jolma, L.F. Campitelli, P.K. Das, Y. Yin, M. Albu, et al., The human transcription factors, *Cell* 172 (4) (2018) 650–665.
- [3] E. Bi, J. Lutkenhaus, FtsZ ring structure associated with division in *Escherichia coli*, *Nature* 354 (6349) (1991) 161–164.
- [4] E. Fuchs, D.W. Cleveland, A structural scaffolding of intermediate filaments in health and disease, *Science* 279 (5350) (1998) 514–519.
- [5] D.W. Nebert, D.W. Russell, Clinical importance of the cytochromes P450, *Lancet* 360 (9340) (2002) 1155–1162.
- [6] J.N. Onuchic, P.G. Wolynes, Theory of protein folding, *Curr. Opin. Struct. Biol.* 14 (1) (2004) 70–75.
- [7] V. Daggett, A. Fersht, The present view of the mechanism of protein folding, *Nat. Rev. Mol. Cell Biol.* 4 (6) (2003) 497–502.
- [8] V. Grantcharova, E.J. Alm, D. Baker, A.L. Horwich, Mechanisms of protein folding, *Curr. Opin. Struct. Biol.* 11 (1) (2001) 70–82.
- [9] D. Baker, A surprising simplicity to protein folding, *Nature* 405 (6782) (2000) 39–42.
- [10] S. Gianni, N.R. Guydosh, F. Khan, T.D. Caldas, U. Mayor, G.W. White, et al., Unifying features in protein-folding mechanisms, *Proc. Natl. Acad. Sci.* 100 (23) (2003) 13286–13291.
- [11] F. Jacob, Evolution and tinkering, *Science* 196 (4295) (1977) 1161–1166.

- [12] S. Ohno, Evolution by Gene Duplication, Springer Science & Business Media, 2013.
- [13] M.T. Levine, C.D. Jones, A.D. Kern, H.A. Lindfors, D.J. Begun, Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression, *Proc. Natl. Acad. Sci.* 103 (26) (2006) 9935–9939.
- [14] D.J. Begun, H.A. Lindfors, M.E. Thompson, A.K. Holloway, Recently evolved genes identified from *Drosophila yakuba* and *D. erecta* accessory gland expressed sequence tags, *Genetics* 172 (3) (2006) 1675–1681.
- [15] Bungard D, Copple JS, Yan J, Chhun JJ, Kumirov VK, Foy SG, et al. Foldability of a natural de novo evolved protein. *Structure* 2017;25(11):1687–96. e4.
- [16] A. Lange, P.H. Patel, B. Heames, A.M. Damry, T. Saenger, C.J. Jackson, et al., Structural and functional characterization of a putative de novo gene in *Drosophila*, *Nat. Commun.* 12 (1) (2021) 1667.
- [17] A.-R. Carvunis, T. Rolland, I. Wapinski, M.A. Calderwood, M.A. Yildirim, N. Simonis, et al., Proto-genes and de novo gene birth, *Nature* 487 (7407) (2012) 370–374.
- [18] H.T. Baalsrud, O.K. Tørresen, M.H. Solbakken, W. Salzburger, R. Hanel, K. S. Jakobsen, et al., De novo gene evolution of antifreeze glycoproteins in codfishes revealed by whole genome sequence data, *Mol. Biol. Evol.* 35 (3) (2018) 593–606.
- [19] A.R. Davidson, K.J. Lumb, R.T. Sauer, Cooperatively folded proteins in random sequence libraries, *Nat. Struct. Biol.* 2 (10) (1995) 856–864.
- [20] A.D. Keefe, J.W. Szostak, Functional proteins from a random-sequence library, *Nature* 410 (6829) (2001) 715–718.
- [21] Protein Data Bank, The single global archive for 3D macromolecular structure data, *Nucleic Acids Res.* 47 (D1) (2019) D520–D528.
- [22] H. Berman, K. Henrick, H. Nakamura, Announcing the worldwide protein data bank, *Nat. Struct. Mol. Biol.* 10 (12) (2003) 980.
- [23] Consortium U, UniProt: a worldwide hub of protein knowledge, *Nucleic Acids Res.* 47 (D1) (2019) D506–D515.
- [24] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, et al., Highly accurate protein structure prediction with AlphaFold, *Nature* 596 (7873) (2021) 583–589.
- [25] J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, et al., Accurate structure prediction of biomolecular interactions with AlphaFold 3, *Nature* 630 (8016) (2024) 493–500.
- [26] R. Krishna, J. Wang, W. Ahern, P. Sturmfels, P. Venkatesh, I. Kalvet, et al., Generalized biomolecular modeling and design with RoseTTAFold all-atom, *Science* 384 (6693) (2024) ead12528.
- [27] A. Lauko, S.J. Pellock, K.H. Sumida, I. Anishchenko, D. Juergens, W. Ahern, et al., Computational design of serine hydrolases, *Science* 388 (6744) (2025) eadu2454.
- [28] D. Röthlisberger, O. Khersonsky, A.M. Wollacott, L. Jiang, J. DeChancie, J. Betker, et al., Kemp elimination catalysts by computational enzyme design, *Nature* 453 (7192) (2008) 190–195.
- [29] L. Jiang, E.A. Althoff, F.R. Clemente, L. Doyle, D. Rothlisberger, A. Zanghellini, et al., De novo computational design of retro-aldol enzymes, *Science* 319 (5868) (2008) 1387–1391.
- [30] S. Vázquez Torres, M. Benard Valle, S.P. Mackessy, S.K. Menzies, N.R. Casewell, S. Ahmadi, et al., De novo designed proteins neutralize lethal snake venom toxins, *Nature* 1–7 (2025).
- [31] L. Qi, E. Bennett, M. Isalan, A directed evolution protocol for engineering minimal transcription factors, based on CIS display, in: *Mammalian Synthetic Systems*, Springer, 2024, pp. 1–13.
- [32] C.B. Anfinsen, Principles that govern the folding of protein chains, *Science* 181 (4096) (1973) 223–230.
- [33] T. Blundell, D. Carney, S. Gardner, F. Hayes, B. Howlin, T. Hubbard, et al., Knowledge-based protein modelling and design, *Eur. J. Biochem.* 172 (3) (1988) 513–520.
- [34] J.U. Bowie, R. Lüthy, D. Eisenberg, A method to identify protein sequences that fold into a known three-dimensional structure, *Science* 253 (5016) (1991) 164–170.
- [35] D.T. Jones, W. Taylor, J.M. Thornton, A new approach to protein fold recognition, *Nature* 358 (6381) (1992) 86–89.
- [36] K.U. Linderstrøm-Lang, *Lane Medical Lectures: Proteins and Enzymes*, Stanford University Press, 1952.
- [37] A.V. Guzzo, The influence of amino acid sequence on protein structure, *Biophys. J.* 5 (6) (1965) 809–822.
- [38] C. Chothia, A.M. Lesk, The relation between the divergence of sequence and structure in proteins, *EMBO J.* 5 (4) (1986) 823–826.
- [39] A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, et al., SWISS-MODEL: homology modelling of protein structures and complexes, *Nucleic Acids Res.* 46 (W1) (2018) W296–W303.
- [40] J. Ye, S. McGinnis, T.L. Madden, BLAST: improvements for better sequence analysis, *Nucleic Acids Res.* 34 (suppl_2) (2006) W6–W9.
- [41] S.R. Eddy, Accelerated profile HMM searches, *PLoS Comput. Biol.* 7 (10) (2011) e1002195.
- [42] F. Sievers, D.G. Higgins, Clustal omega, *Curr. Protoc. Bioinformatics* 48(1):3.13 (2014) 1–3.. 6.
- [43] D. Baker, A. Sali, Protein structure prediction and structural genomics, *Science* 294 (5540) (2001) 93–96.
- [44] A.A. Ivanov, D. Barak, K.A. Jacobson, Evaluation of homology modeling of G-protein-coupled receptors in light of the A2A adenosine receptor crystallographic structure, *J. Med. Chem.* 52 (10) (2009) 3284–3292.
- [45] A. Andreeva, E. Kulesha, J. Gough, A.G. Murzin, The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures, *Nucleic Acids Res.* 48 (D1) (2020) D376–D382.
- [46] Z. Zhang, S. Kochhar, M.G. Grigorov, Descriptor-based protein remote homology identification, *Protein Sci.* 14 (2) (2005) 431–444.
- [47] R.B. Russell, M. Saqi, P.A. Bates, R.A. Sayle, M. Sternberg, Recognition of analogous and homologous protein folds—assessment of prediction success and associated alignment accuracy using empirical substitution matrices, *Protein Eng.* 11 (1) (1998) 1–9.
- [48] L.A. Kelley, S. Mezulis, C.M. Yates, M.N. Wass, M.J. Sternberg, The Phyre2 web portal for protein modeling, prediction and analysis, *Nat. Protoc.* 10 (6) (2015) 845–858.
- [49] CASP9. <https://www.predictioncenter.org/casp9/index.cgi>, 2010. Accessed 1 July 2025.
- [50] P. Mitra, D. Shultis, Y. Zhang, EvoDesign: de novo protein design based on structural and evolutionary profiles, *Nucleic Acids Res.* 41 (W1) (2013) W273–W280.
- [51] E. Ong, X. Huang, R. Pearce, Y. Zhang, Y. He, Computational design of SARS-CoV-2 spike glycoproteins to increase immunogenicity by T cell epitope engineering, *Comput. Struct. Biotechnol. J.* 19 (2021) 518–529.
- [52] B. Kuhlman, G. Dantas, G.C. Ireton, G. Varani, B.L. Stoddard, D. Baker, Design of a novel globular protein fold with atomic-level accuracy, *Science* 302 (5649) (2003) 1364–1368.
- [53] P.-S. Huang, G. Oberdorfer, C. Xu, X.Y. Pei, B.L. Nannenga, J.M. Rogers, et al., High thermodynamic stability of parametrically designed helical bundles, *Science* 346 (6208) (2014) 481–485.
- [54] H. Shen, J.A. Fallas, E. Lynch, W. Sheffler, B. Parry, N. Jannetty, et al., De novo design of self-assembling helical protein filaments, *Science* 362 (6415) (2018) 705–709.
- [55] A.A. Vorobieva, P. White, B. Liang, J.E. Horne, A.K. Bera, C.M. Chow, et al., De novo design of transmembrane β barrels, *Science* 371 (6531) (2021) eabc8182.
- [56] P.-S. Huang, K. Feldmeier, F. Parmeggiani, D.A. Fernandez Velasco, B. Höcker, D. Baker, De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy, *Nat. Chem. Biol.* 12 (1) (2016) 29–34.
- [57] K.T. Simons, R. Bonneau, I. Ruczinski, D. Baker, Ab initio protein structure prediction of CASP III targets using ROSETTA, *Proteins Struct. Funct. Bioinform.* 37 (S3) (1999) 171–176.
- [58] Protein Structure Prediction Center and Critical Assessment of Techniques for Protein Structure Prediction. <https://www.predictioncenter.org/index.cgi>, 2025. Accessed 1 July 2025.
- [59] A.R. Thomson, C.W. Wood, A.J. Burton, G.J. Bartlett, R.B. Sessions, R.L. Brady, et al., Computational design of water-soluble α -helical barrels, *Science* 346 (6208) (2014) 485–488.
- [60] E. Marcos, T.M. Chidyausiku, A.C. McShan, T. Evangelidis, S. Nerli, L. Carter, et al., De novo design of a non-local β -sheet protein with high stability and accuracy, *Nat. Struct. Mol. Biol.* 25 (11) (2018) 1028–1034.
- [61] J.M. Fletcher, R.L. Harniman, F.R. Barnes, A.L. Boyle, A. Collins, J. Mantell, et al., Self-assembling cages from coiled-coil peptide modules, *Science* 340 (6132) (2013) 595–599.
- [62] J. Dou, A.A. Vorobieva, W. Sheffler, L.A. Doyle, H. Park, M.J. Bick, et al., De novo design of a fluorescence-activating β -barrel, *Nature* 561 (7724) (2018) 485–491.
- [63] M. AlQuraishi, AlphaFold at CASP13, *Bioinformatics* 35 (22) (2019) 4862–4865.
- [64] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G.R. Lee, et al., Accurate prediction of protein structures and interactions using a three-track neural network, *Science* 373 (6557) (2021) 871–876.
- [65] J.L. Watson, D. Juergens, N.R. Bennett, B.L. Trippe, J. Yim, H.E. Eisenach, et al., De novo design of protein structure and function with RFdiffusion, *Nature* 620 (7976) (2023) 1089–1100.
- [66] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, *Adv. Neural Inf. Process. Syst.* 33 (2020) 6840–6851.
- [67] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, et al., Attention is all you need, *Adv. Neural Inf. Process. Syst.* 30 (2017).
- [68] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, *Adv. Neural Inf. Process. Syst.* 25 (2012).
- [69] W. Rawat, Z. Wang, Deep convolutional neural networks for image classification: a comprehensive review, *Neural Comput.* 29 (9) (2017) 2352–2449.
- [70] Q. Li, W. Cai, X. Wang, Y. Zhou, D.D. Feng, M. Chen, Medical image classification with convolutional neural network, in: 2014 13th International Conference on Control Automation Robotics & Vision (ICARCV), IEEE, 2014, pp. 844–848.
- [71] S.S. Yadav, S.M. Jadhav, Deep convolutional neural network based medical image classification for disease diagnosis, *J. Big Data* 6 (1) (2019) 1–18.
- [72] L. Nanni, S. Ghidoni, S. Brahnam, Ensemble of convolutional neural networks for bioimage classification, *Appl. Comput. Inform.* 17 (1) (2021) 19–35.
- [73] G.-W. Wei, Protein structure prediction beyond AlphaFold, *Nat. Mach. Intell.* 1 (8) (2019) 336–337.
- [74] CASP13. <https://www.predictioncenter.org/casp13/index.cgi>, 2018. Accessed 1 July 2025.
- [75] W.D. Heaven, Deepmind’s protein-folding AI has solved a 50-year-old grand challenge of biology, *MIT, Technol. Rev.* (2020) 1012712.
- [76] Service RF, ‘The Game Has Changed.’ AI Triumphs at Protein Folding, American Association for the Advancement of Science, 2020.
- [77] C.J. Glasscock, R. Pecoraro, R. McHugh, L.A. Doyle, W. Chen, O. Boivin, et al., Computational design of sequence-specific DNA-binding proteins, *Nat. Struct. Mol. Biol.* (2025) 1–10.

- [78] D.E. Kim, D.R. Jensen, D. Feldman, D. Tischer, A. Saleem, C.M. Chow, et al., De novo design of small beta barrel proteins, *Proc. Natl. Acad. Sci.* 120 (11) (2023) e2207974120.
- [79] D.R. Hicks, M.A. Kennedy, K.A. Thompson, M. DeWitt, B. Coventry, A. Kang, et al., De novo design of protein homodimers containing tunable symmetric protein pockets, *Proc. Natl. Acad. Sci.* 119 (30) (2022) e2113400119.
- [80] D.D. Saitoe, F. Praetorius, A. Courbet, Y. Hsia, B.I. Wicky, N.I. Edman, et al., Reconfigurable asymmetric protein assemblies through implicit negative design, *Science* 375 (6578) (2022) eabj7662.
- [81] L. Cao, B. Coventry, I. Goresnik, B. Huang, W. Sheffler, J.S. Park, et al., Design of protein-binding proteins from the target structure alone, *Nature* 605 (7910) (2022) 551–560.
- [82] W. Zheng, Q. Wuyun, Y. Li, Q. Liu, X. Zhou, C. Peng, et al., Deep-learning-based single-domain and multidomain protein structure prediction with DI-TASSER, *Nat. Biotechnol.* 1–13 (2025).
- [83] W. Zheng, Q. Wuyun, Y. Li, C. Zhang, L. Freddolino, Y. Zhang, Improving deep learning protein monomer and complex structure prediction using DeepMSA2 with huge metagenomics data, *Nat. Methods* 21 (2) (2024) 279–289.
- [84] Y. Li, C. Zhang, D.-J. Yu, Y. Zhang, Deep learning geometrical potential for high-accuracy ab initio protein structure prediction, *Science* 25 (6) (2022).
- [85] W. Zheng, Q. Wuyun, X. Zhou, Y. Li, L. Freddolino, Y. Zhang, LOMETS3: integrating deep learning and profile alignment for advanced protein template recognition and function annotation, *Nucleic Acids Res.* 50 (W1) (2022) W454–W464.
- [86] R.H. Swendsen, J.-S. Wang, Replica Monte Carlo simulation of spin-glasses, *Phys. Rev. Lett.* 57 (21) (1986) 2607.
- [87] CASP15. <https://predictioncenter.org/casp15/>, 2022. Accessed 22nd of Aug 2025.
- [88] C.A. Rohl, C.E. Strauss, K.M. Misura, D. Baker, Protein structure prediction using Rosetta, in: *Methods in Enzymology*, Elsevier, 2004, pp. 66–93.
- [89] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, et al., Evolutionary-scale prediction of atomic-level protein structure with a language model, *Science* 379 (6637) (2023) 1123–1130.
- [90] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, L. Van Gool, Repaint: Inpainting using denoising diffusion probabilistic models, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11461–11471.
- [91] K.E. Wu, K.K. Yang, R. van den Berg, S. Alamdari, J.Y. Zou, A.X. Lu, et al., Protein structure generation via folding diffusion, *Nat. Commun.* 15 (1) (2024) 1059.
- [92] S.L. Lisanza, J.M. Gershon, S.W. Tipps, W.N. Sims, L. Arnoldt, S.J. Hendel, et al., Multistate and functional protein design using RoseTTAFold sequence space diffusion, *Nat. Biotechnol.* 1–11 (2024).
- [93] A.E. Chu, J. Kim, L. Cheng, G. El Nesr, M. Xu, R.W. Shuai, et al., An all-atom protein generative model, *Proc. Natl. Acad. Sci.* 121 (27) (2024) e2311500121.
- [94] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R.J. Ragotte, L.F. Milles, et al., Robust deep learning-based protein sequence design using ProteinMPNN, *Science* 378 (6615) (2022) 49–56.
- [95] T. Yu, H. Cui, J.C. Li, Y. Luo, G. Jiang, H. Zhao, Enzyme function prediction using contrastive learning, *Science* 379 (6639) (2023) 1358–1363.
- [96] J. Tubiana, D. Schneidman-Duhovny, H.J. Wolfson, ScanNet: an interpretable geometric deep learning model for structure-based protein binding site prediction, *Nat. Methods* 19 (6) (2022) 730–739.
- [97] J. Wohlwend, G. Corso, S. Passaro, M. Reveiz, K. Leidal, W. Swiderski, et al., Boltz-1 democratizing biomolecular interaction modeling, *BioRxiv* (2024), <https://doi.org/10.1101/2024.11.19.624167>.
- [98] S. Passaro, G. Corso, J. Wohlwend, M. Reveiz, S. Thaler, V.R. Somnath, et al., Boltz-2: towards accurate and efficient binding affinity prediction, *BioRxiv* 2025 (06) (2025), 14.659707.
- [99] J. Hirschberg, C.D. Manning, Advances in natural language processing, *Science* 349 (6245) (2015) 261–266.
- [100] K. Chowdhary, K. Chowdhary, Natural language processing, *Fundam. Artif. Intell.* (2020) 603–649.
- [101] N. Ferruz, S. Schmidt, B. Höcker, ProtGPT2 is a deep unsupervised language model for protein design, *Nat. Commun.* 13 (1) (2022) 4348.
- [102] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1 (Long and Short Papers), 2019, pp. 4171–4186.
- [103] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, Y. Liu, Roformer: enhanced transformer with rotary position embedding, *Neurocomputing* 568 (2024) 127063.
- [104] Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, Consortium U, UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches, *Bioinformatics* 31 (6) (2015) 926–932.
- [105] B.E. Suzek, H. Huang, P. McGarvey, R. Mazumder, C.H. Wu, UniRef: comprehensive and non-redundant UniProt reference clusters, *Bioinformatics* 23 (10) (2007) 1282–1288.
- [106] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, et al., Gene ontology: tool for the unification of biology, *Nat. Genet.* 25 (1) (2000) 25–29.
- [107] I. Anishchenko, S.J. Pellock, T.M. Chidyausiku, T.A. Ramelot, S. Ovchinnikov, J. Hao, et al., De novo protein design by deep network hallucination, *Nature* 600 (7889) (2021) 547–552.
- [108] B. Wicky, L. Milles, A. Courbet, R. Ragotte, J. Dauparas, E. Kinfu, et al., Hallucinating symmetric protein assemblies, *Science* 378 (6615) (2022) 56–61.
- [109] Hie B, Candido S, Lin Z, Kabeli O, Rao R, Smetanin N, et al. A high-level programming language for generative protein design. *BioRxiv* 2022:2022.12.21.521526.
- [110] Verkuil R, Kabeli O, Du Y, Wicky BI, Milles LF, Dauparas J, et al. Language models generalize beyond natural proteins. *BioRxiv* 2022:2022.12.21.521521.
- [111] S. Vázquez Torres, P.J. Leung, P. Venkatesh, I.D. Lutz, F. Hink, H.-H. Huynh, et al., De novo design of high-affinity binders of bioactive helical peptides, *Nature* 626 (7998) (2024) 435–442.
- [112] S. Lyskov, J.J. Gray, The RosettaDock server for local protein–protein docking, *Nucleic Acids Res.* 36 (suppl_2) (2008) W233–W238.
- [113] G.M. Morris, R. Huey, A.J. Olson, Using autodock for ligand-receptor docking, *Curr. Protoc. Bioinformatics* 24 (1) (2008) 8, 14. 1–8. 40.
- [114] S.A. Hollingsworth, R.O. Dror, Molecular dynamics simulation for all, *Neuron* 99 (6) (2018) 1129–1143.
- [115] M. Oeller, R. Kang, R. Bell, H. Ausserwöger, P. Sormanni, M. Vendruscolo, Sequence-based prediction of pH-dependent protein solubility using CamSol, *Brief. Bioinform.* 24 (2) (2023) bbad004.
- [116] S. Khurana, R. Rawi, K. Kunji, G.-Y. Chuang, H. Bensmail, R. Mall, DeepSol: a deep learning framework for sequence-based protein solubility prediction, *Bioinformatics* 34 (15) (2018) 2605–2613.
- [117] V. Thumhuri, H.-M. Martiny, J.J. Almagro Armenteros, J. Salomon, H. Nielsen, A.R. Johansen, NetSolP: predicting protein solubility in *Escherichia coli* using language models, *Bioinformatics* 38 (4) (2022) 941–946.
- [118] T.-E. Kim, K. Tsuboyama, S. Houlston, C.M. Martell, C.M. Phoumyvong, A. Lemak, et al., Dissecting the stability determinants of a challenging de novo protein fold using massively parallel design and experimentation, *Proc. Natl. Acad. Sci.* 119 (41) (2022) e2122676119.
- [119] H. Kaessmann, Origins, evolution, and phenotypic impact of new genes, *Genome Res.* 20 (10) (2010) 1313–1326.
- [120] W.R. Blevins, J. Ruiz-Orera, X. Messegue, B. Blasco-Moreno, J.L. Villanueva-Cañas, L. Espinar, et al., Uncovering de novo gene birth in yeast using deep transcriptomics, *Nat. Commun.* 12 (1) (2021) 604.
- [121] P.-P. Grassé, Evolution of Living Organisms: Evidence for a New Theory of Transformation, Academic Press, 2013.
- [122] J. Cai, R. Zhao, H. Jiang, W. Wang, De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae*, *Genetics* 179 (1) (2008) 487–496.
- [123] A.M. Gubala, J.F. Schmitz, M.J. Kearns, T.T. Vinh, E. Bornberg-Bauer, M. F. Wolfner, et al., The Goddard and sartum genes are essential for *Drosophila* male fertility and may have arisen de novo, *Mol. Biol. Evol.* 34 (5) (2017) 1066–1082.
- [124] X. Zhuang, C. Yang, K.R. Murphy, C.-H.C. Cheng, Molecular mechanism and history of non-sense to sense evolution of antifreeze glycoprotein gene in northern gadids, *Proc. Natl. Acad. Sci.* 116 (10) (2019) 4400–4405.
- [125] L. Li, C.M. Foster, Q. Gan, D. Nettleton, M.G. James, A.M. Myers, et al., Identification of the novel protein QQS as a component of the starch metabolic network in *Arabidopsis* leaves, *Plant J.* 58 (3) (2009) 485–498.
- [126] C. Xie, C. Bekpen, S. Künzel, M. Keshavarz, R. Krebs-Wheaton, N. Skrabar, et al., A de novo evolved gene in the house mouse regulates female pregnancy cycles, *Elife* 8 (2019) e44392.
- [127] D.G. Knowles, A. McLysaght, Recent de novo origin of human protein-coding genes, *Genome Res.* 19 (10) (2009) 1752–1759.
- [128] J. Peng, L. Zhao, The origin and structural evolution of de novo genes in *Drosophila*, *Nat. Commun.* 15 (1) (2024) 810.
- [129] V. Mariani, M. Biasini, A. Barbatto, T. Schwede, IDDT: a local superposition-free score for comparing protein structures and models using distance difference tests, *Bioinformatics* 29 (21) (2013) 2722–2728.
- [130] Y. Zhang, J. Skolnick, Scoring function for automated assessment of protein structure template quality, *Proteins: Struct., Funct., Bioinf.* 57 (4) (2004) 702–710.
- [131] C. Chiarabelli, J.W. Vrijbloed, R.M. Thomas, P.L. Luisi, Investigation of de novo totally random biosequences, part I: a general method for in vitro selection of folded domains from a random polypeptide library displayed on phage, *Chem. Biodivers.* 3 (8) (2006) 827–839.
- [132] C. Chiarabelli, J.W. Vrijbloed, D. De Luca, R.M. Thomas, P. Stano, F. Polticelli, et al., Investigation of de novo totally random biosequences, part II: on the folding frequency in a totally random library of de novo proteins obtained by phage display, *Chem. Biodivers.* 3 (8) (2006) 840–859.
- [133] I.D. Prijambada, T. Yomo, F. Tanaka, T. Kawama, K. Yamamoto, A. Hasegawa, et al., Solubility of artificial proteins with random sequences, *FEBS Lett.* 382 (1–2) (1996) 21–25.
- [134] J. Tanaka, N. Doi, H. Takashima, H. Yanagawa, Comparative characterization of random-sequence proteins consisting of 5, 12, and 20 kinds of amino acids, *Protein Sci.* 19 (4) (2010) 786–795.
- [135] M.S. Newton, D.J. Morrone, K.H. Lee, B. Seelig, Genetic code evolution investigated through the synthesis and characterisation of proteins from reduced-alphabet libraries, *ChemBioChem* 20 (6) (2019) 846–856.
- [136] S.S. Mansy, J. Zhang, R. Kümmerle, M. Nilsson, J.J. Chou, J.W. Szostak, et al., Structure and evolutionary analysis of a non-biological ATP-binding protein, *J. Mol. Biol.* 371 (2) (2007) 501–513.
- [137] J.C. Chaput, J.W. Szostak, Evolutionary optimization of a nonbiological ATP binding protein for improved folding stability, *Chem. Biol.* 11 (6) (2004) 865–874.
- [138] P.L. Surdo, M.A. Walsh, M. Sollazzo, A novel ADP-and zinc-binding fold from function-directed in vitro evolution, *Nat. Struct. Mol. Biol.* 11 (4) (2004) 382–383.

- [139] J.-F. Yu, Z. Cao, Y. Yang, C.-L. Wang, Z.-D. Su, Y.-W. Zhao, et al., Natural protein sequences are more intrinsically disordered than random sequences, *Cell. Mol. Life Sci.* 73 (2016) 2949–2957.
- [140] A.F. Ángyán, A. Perczel, Z. Gáspári, Estimating intrinsic structural preferences of de novo emerging random-sequence proteins: is aggregation the main bottleneck? *FEBS Lett.* 586 (16) (2012) 2468–2472.
- [141] V. Tretyachenko, J. Vymětal, L. Bednářová, V. Kopecký Jr., K. Hofbauerová, H. Jindrová, et al., Random protein sequences can form defined secondary structures and are well-tolerated in vivo, *Sci. Rep.* 7 (1) (2017) 15449.
- [142] G. Cho, A.D. Keefe, R. Liu, D.S. Wilson, J.W. Szostak, Constructing high complexity synthetic libraries of long ORFs using in vitro selection, *J. Mol. Biol.* 297 (2) (2000) 309–319.
- [143] D.S. Wilson, A.D. Keefe, J.W. Szostak, The use of mRNA display to select high-affinity protein-binding peptides, *Proc. Natl. Acad. Sci.* 98 (7) (2001) 3750–3755.
- [144] T.T. Takahashi, R.J. Austin, R.W. Roberts, mRNA display: ligand discovery, interaction analysis and beyond, *Trends Biochem. Sci.* 28 (3) (2003) 159–165.
- [145] B. Seelig, mRNA display for the selection and evolution of enzymes from in vitro-translated protein libraries, *Nat. Protoc.* 6 (4) (2011) 540–552.
- [146] R. Odegrip, D. Coomber, B. Eldridge, R. Hederer, P.A. Kuhlman, C. Ullman, et al., CIS display: in vitro selection of peptides from libraries of protein–DNA complexes, *Proc. Natl. Acad. Sci.* 101 (9) (2004) 2806–2810.
- [147] B. Cummins, J. Vrana, R.C. Moseley, H. Eramian, A. Deckard, P. Fontanarrosa, et al., Robustness and reproducibility of simple and complex synthetic logic circuit designs using a DBTL loop, *Synth. Biol.* 8 (1) (2023) ysad005.
- [148] A.K. Brödel, R. Rodrigues, A. Jaramillo, M. Isalan, Accelerated evolution of a minimal 63-amino acid dual transcription factor, *Sci. Adv.* 6 (24) (2020) eaba2728.
- [149] S.M. Miller, T. Wang, D.R. Liu, Phage-assisted continuous and non-continuous evolution, *Nat. Protoc.* 15 (12) (2020) 4101–4127.
- [150] R. Chao, S. Mishra, T. Si, H. Zhao, Engineering biological systems using automated biofoundries, *Metab. Eng.* 42 (2017) 98–108.
- [151] Holowko MB, Frow EK, Reid JC, Rourke M, Vickers CE. Building a biofoundry. *Synth. Biol.* 2021;6(1):ysaa026.