Generative Models in Protein Engineering: A Comprehensive Survey

Xinhui Chen[1*] **Yiwen Yuan**[3*] **Joseph Liu**[4*]
Chak Tou Leong[5] **Xiaoye Zhu**[6] **Jiaqi Chen**[2†]

[1]Wuhan University [2]Fudan University [3]Carnegie Mellon University
[4]University of Southern California [5]Hong Kong Polytechnic University
[6]National University of Singapore

https://github.com/XinhuiChen-02/Generative-Protein-Modeling

Abstract

Proteins are fundamental molecules performing diverse functions in living organisms. Protein engineering, the process of designing or modifying proteins to enhance or create new functions, has therefore become a research focus in the fields of biotechnology and medicine. A primary challenge in protein engineering is to efficiently discover and design new proteins with desired functions. Traditional approaches like directed evolution and rational design, though widely used, are limited by high computational costs and restricted exploration of potential protein structures. The recent success of generative models in efficiently synthesizing high-quality data across various domains has inspired researchers to investigate their potential applications in protein engineering. In this survey, we systematically summarize recent works on generative models for protein engineering, with a particular focus on protein design. Specifically, we categorize three main frameworks in existing generative protein design methods: sequence-based, structure-based, and joint sequence-structure generation. Besides, we provide a detailed review of representative generative models, including autoregressive models and diffusion models, and their application in protein sequence prediction and structure generation. Finally, we pinpoint existing challenges and propose future directions, such as leveraging large datasets, improving complex structure validation, and integrating advanced modeling techniques.

1 Introduction

Proteins are fundamental molecules that perform a wide range of functions in living organisms, exhibiting diverse structures and functionalities. They are involved in processes such as biological catalysis, signal transduction, and structural support. Due to their crucial roles in catalysis and functional regulation, ***protein engineering***, the process of designing or modifying proteins to enhance existing functions or create new ones, plays a vital role in biotechnology and medicine. For example, in biotechnology, industrially engineered enzymes are widely used in detergents, enhancing their stability under high temperature and alkaline conditions [42]. In the medical field, pembrolizumab (Keytruda) significantly improves treatment outcomes for various cancers by targeting the immune

*These authors contributed equally to this work.
†Corresponding author.

Foundation Models for Science Workshop, 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

checkpoint PD-1, effectively boosting patients' immune responses [57, 63, 44]. Additionally, protein engineering is used in developing new drugs, designing biomaterials, and environmental remediation [38]. However, a major challenge in protein engineering is efficiently discovering and designing novel proteins with specific desired functions, namely *protein design*. Traditional approaches like directed evolution and rational design have been instrumental in protein engineering [2, 74]. However, these methods are often time-consuming, costly, and offer limited exploration of protein variants [74]. Directed evolution requires the construction and screening of large mutant libraries, while rational design is constrained by our incomplete understanding of protein structure-function relationships.

Machine learning approaches, particularly deep learning, now enable efficient exploration of vast biological data and accurate prediction of molecular properties [13]. Generative models, a subset of deep learning, have proven highly effective in creating and evaluating novel protein sequences in protein engineering [68]. By encoding the protein sequences into embedding space and training with large-scale datasets of known protein sequences, these models can learn the complex sequence-structure-function relationships efficiently. Once trained, the models can generate and preliminarily screen vast numbers of candidate sequences, significantly accelerating the protein discovery process [49, 68]. In addition, generative models can also capture subtle patterns that human experts might overlook, offering new insights for protein design [40]. For instance, generative models incorporating AlphaFold2's structural prediction techniques can leverage its ability to model long-range dependencies in amino acid sequences, identifying complex interactions like hydrogen bonds and hydrophobic effects, thereby enhancing the design of novel functional proteins [36, 90]. In summary, generative models are revolutionizing the field of protein engineering, opening up new possibilities for solving complex protein design challenges.

In this survey, we present a mind map that systematically categorizes and analyzes the problem definitions and existing methods for protein design, providing a comprehensive overview. Our major contributions are as follows:

- We present a systematic categorization of protein design approaches, focusing on three main frameworks: sequence-based, structure-based, and joint sequence-structure generation in Section 2.
- We provide a focused review of two primary types of generative models in Section 3: autoregressive models and diffusion models, discussing their principles, architectures, and applications in protein sequence prediction and structure generation.
- We discuss the challenges and opportunities in generative protein models, identifying open problems and future research directions in Section 4, to pave the way for the next generation of protein engineering.

2 Sequence, Structure, and Joint Generation

In this section, we categorize common generative tasks by input and output in the area of protein engineering. For each task, we also summarize the major model architectures in solving different tasks.

2.1 Sequence-based Protein Design

In sequence-based protein sequence generation tasks, the models learn the evolutionary relationship of protein sequences directly from multidimensional amino-acid sequence space, with or without functional or structural constraints. The goal is to generate new sequences that possess desired properties while meeting the given constraints. Autoregressive transformer-based models [19, 47, 52, 22] have shown excellent performance in this task, capable of learning long-range dependencies and complex sequence patterns.

Probabilistic graphical models (PGMs) represent the conditional probability distribution over random variables. The structure of graphs provides efficiency through variable elimination. It has been broadly applied in structure prediction and alignment tasks to predict the conditional probability of the structure conditional on the given sequences [30]. Nevertheless, PGMs have also been used in generating new protein sequences [70, 56].

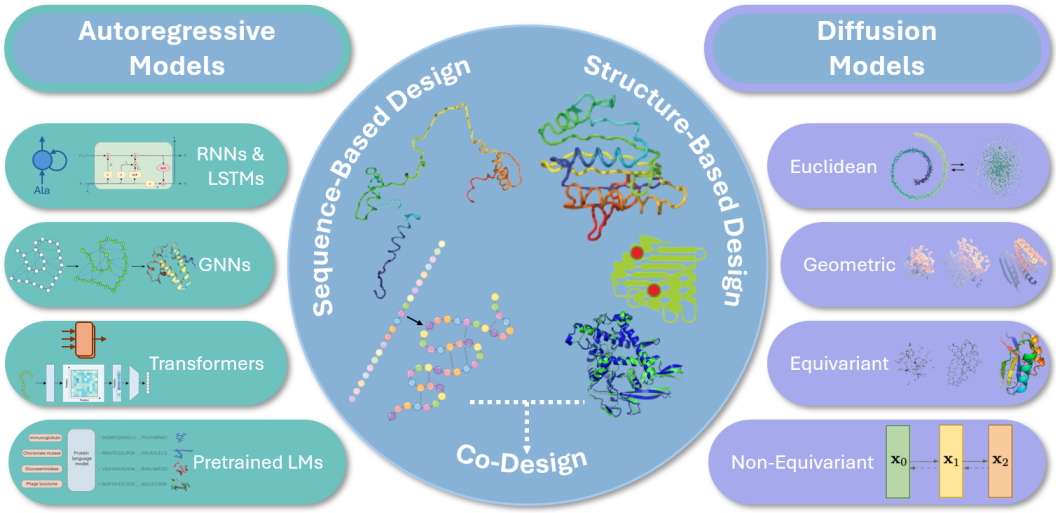

Figure 1: **Generative models in protein design.** Overview of autoregressive (left) and diffusion (right) generative models in protein design. The central circle depicts core sequence-based and structure-based design tasks.

2.2 Structure-based Protein Design

Another common task in protein engineering is to predict sequences from structural information (Figure 2). Graph Neural Networks (GNN) is a common architecture on tasks with structural information as input as it can perform both spatial and relational reasoning. GVP-GNN [33] can be applied to any problem where the input domain is a structure of a single macromolecule or of molecules bound to one another. Other approaches focus more on specific task domains. Recently developed GPSFun [87] demonstrates how geometric GNNs can capture both sequence and structural patterns from protein graphs to enable comprehensive function predictions ranging from binding sites to gene ontologies and subcellular locations. Diffusion models [20, 89, 71] and PGMs[41] have both been applied to "inpaint" missing protein sequence segments given known structures. Message-Passing Neural Networks (MPNN) [17] and Transformers [81] have been applied to generating full-sequence given structural input.

2.3 Sequence and Structure Co-generation

The co-generation of protein sequences and structures is crucial because it ensures their compatibility, which is essential for both functional and structural correctness. By designing sequence-structure data pairs as output, researchers can prevent mismatches that arise when changes in the sequence lead to incompatible structures. This approach is especially important in fields like de novo protein design and drug discovery, where precise control of structure and function is critical. Diffusion models [62, 66, 77] and Graph Neural Networks (GNN) [31], and their combination [29] have been adopted for solving this task.

3 Generative methods

3.1 Autoregressive models

3.1.1 Recurrent Neural Networks and Long Short-Term Memory

Autoregressive models (ARs) using Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTMs) have shown great potential in protein sequence generation. Early RNNs demonstrate the ability to capture long-range dependencies in sequences [4], but face issues with vanishing/exploding gradients for long sequences [6]. To address this, Bidirectional RNNs (BRNNs) [59] are introduced, improving contextual information capture by processing sequences bidirectionally [61], especially in secondary structure prediction tasks. Deep RNN architectures, in which layers of RNNs are stacked

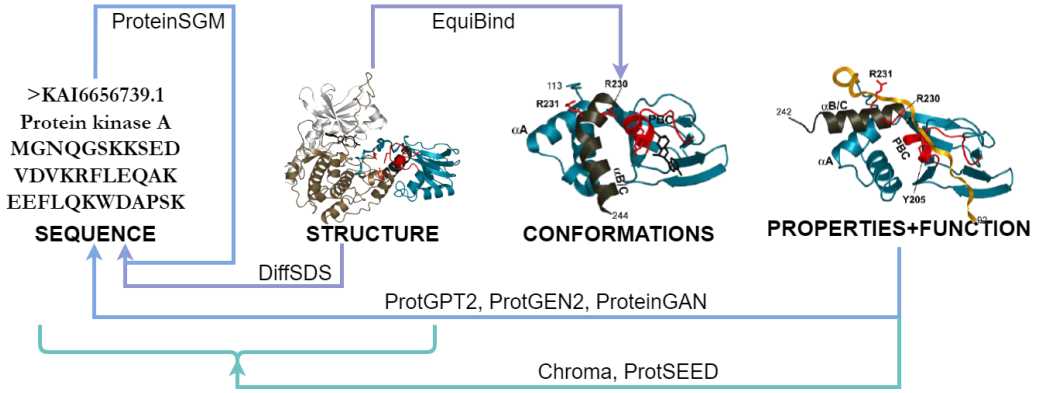

Figure 2: **The interplay between various elements in protein design algorithms.** At its core, proteins are depicted as amino acid chains that adopt three-dimensional configurations, where a single low-energy structure can be distinguished from multiple possible conformations. Structural regulation is represented as the modulation of protein function through chemical modifications. Color-coded arrows represent different algorithmic approaches: blue for sequence-based, purple for structure-based, and green for sequence and structure co-generation methods. This framework outlines the present state and future directions of computational protein design algorithms.

for more abstract feature representation, achieve success in complex tasks like protein function prediction [45]. The CRRNN model further combines CNNs, residual connections, and RNNs for local feature extraction and sequence modeling, leading to state-of-the-art secondary structure prediction [88].

LSTMs [23] address long-range dependency issues by introducing gating mechanisms, enabling improved modeling of complex sequence patterns [64]. Bidirectional LSTMs (BiLSTMs) enhance performance further by processing sequences in both directions [76, 59]. DeepANIS, combining BiLSTM and Transformer networks, significantly improves antibody paratope prediction [91]. Despite advances, challenges remain in computational efficiency and interpretability [36], suggesting future research should focus on integrating sequential models with other advanced techniques to improve prediction accuracy and protein biology understanding. Key characteristics and noteworthy facts about RNN and LSTM models are summarized in Table 1.

3.1.2 Transformers and Pretrained Language Models

The Transformer architecture [72] has proven its value as an effective approach in protein science and engineering. The self-attention mechanism learns the pairwise interactions between all positions in the sequence. Stacking multiple self-attention layers results in learning multi-residual interactions [75]. The positional encoding captures the sequential information. Since the release of the Transformer architecture, researchers have been actively utilizing it in various protein-related tasks. A collection of notable information on Transformers-based autoregressive models in protein design is provided in Table 1. Wu *et al.* [82] mapped signal peptide generation to a machine translation problem. Pre-training on large text corpora has shown promise in large language models on generating human text; the pre-training strategy adapts well to the area of protein generation. Rao *et al.* [55] show that self-supervised pre-training improves overall performance across tasks such as remote homology detection and stability landscape prediction.

Pre-training has proven to be effective in protein generation tasks in either supervised or unsupervised settings. Transformer-based models such as ProtGPT2 [19], pre-trained on 50M unannotated protein sequences, and ProGen2 [47], pre-trained on 280M sequences with metadata tags, are capable of generating sequences in line with natural proteins. Furthermore, ProGen2 [47] also supports controlled generation based on tags representing protein family, biological process, and molecular function. ZymCTRL [51] specifically addresses the problem of artificial enzyme generation. Trained on more than 36 million enzyme sequences from the BRENDA database, ZymCTRL [51] can generate enzyme sequences with user-specified catalytic functions.

Table 1: **Summary of autoregressive models by architecture, input-output relations, and functional properties.** The functional properties column distinguishes whether the model can generate a protein with a specific function. 'Struct.' denotes 'structure', 'Seq.' denotes 'sequence'.

Model type	Method	Input	Output	Functional properties
RNN	Deep RNN (2017) [45]	Protein Seq.	Secondary Struct.	No
	CRRNN (2018) [88]	Protein Seq.	Structural class	No
LSTM	DeepANIS (2021) [91]	Protein Seq.	PPI sites	Yes (antimicrobial)
Transformer	ProtGPT2 (2022) [19]	Prompt/conditioning	Protein Seq.	Yes (task-specific)
	ZymCTRL (2022) [51]	Enzyme properties	Enzyme Seq.	Yes (enzymatic)
	ProteinSGM (2022) [41]	Seq./conditions	Protein Struct.	No
	OmegaFold (2022) [81]	Protein Seq.	Protein 3D Struct.	No
	ProtSEED (2022) [62]	Protein Seq.	Protein Struct.	Yes (task-specific)
	ProGen2 (2023) [47]	Conditioning tags	Protein Seq.	Yes (task-specific)
GNN	GVP-GNN (2020) [33]	3D protein Struct.	Anything	No
	EquiBind (2022) [67]	Protein & ligand Struct.	Binding pose	Yes (binding)
	ProteinMPNN (2022) [17]	Backbone Struct.	Protein Seq.	No
	RefineGNN (2022) [31]	Iterative antibody Seq.	Optimized Seq.	Yes (binding)
	Abode (2023) [73]	Antigen & initial antibody	Optimized antibody	Yes (antibody)

3.1.3 Graph Neural Networks

Message passing is a computationally efficient solution for identifying direct interactions from residue positions in protein sequences, as traditional correlation methods would require pairwise calculations. [79] Message Passing Graph Neural Networks (GNNs) is a natural way to learn representations of proteins, as each protein is a polymer of amino acids. Typically, each residue in a protein is represented as a node in a graph, and the neighborhood of a node is the set of neighboring nodes in the protein structure. The node features and edge features can be used to capture spatial and residue-level information. Combined with an autoregressive decoder, GNNs have been used in the tasks of protein sequence design [29, 17], antibody structure design [31, 73, 37].

Both Ingraham *et al.* and Dauparas *et al.* focused on generating the desired sequence that folds to a desired structure. Ingraham *et al.* [29] considered a graph of the k nearest neighbors with invariant and locally informative edge features. The model has an encoder-decoder architecture and uses self-attention as the aggregation method. Following the graph design of Ingraham *et al.*, Dauparas *et al.* [17] discovered that additional backbone noise and higher inference temperature can improve the sequence recovery rate and the robustness of generated sequences, showing that techniques that facilitate language generation also have a positive influence on sequence generation.

Antibodies are one of the proteins that protect the body from harmful substances. These Y-shaped proteins recognize antigens through their complementarity-determining regions (CDRs), which consist of short peptide sequences. Jin *et al.* [31] modeled the CDR design of antibodies as a graph generation problem. Their proposed RefineGNN [31] autoregressively predicts the next residue based on the node embedding of the last generated residue. Verma *et al.* [73] tackled the task of designing antibodies that target specific antigens. It models the antibody-antigen complex as a 3-D heterogeneous graph and generates antibody sequences by solving a system of ordinary differential equations (ODEs) describing the continuous evolution of node states. Notable information on GNN-based autoregressive models is summarized in Table 1.

3.2 Diffusion models

3.2.1 Geometry in protein structure: euclidean *v.s.* geometric

In the field of protein structure generation, the development of diffusion models has undergone a paradigm shift from Euclidean to geometric approaches, reflecting a deep understanding of the intrinsic properties of protein structures and a qualitative leap in modeling capabilities [85]. The comparison between these two types of methods is provided in Table 2. Euclidean diffusion models

simplify protein structures as point sets in Euclidean space, describing the generation process through the following stochastic differential equation (SDE):

$$dx = [f(x,t) - g(t)^2 \nabla \log p_t(x)]dt + g(t)dB \tag{1}$$

where $f(x,t)$ is the drift term, $g(t)$ is the diffusion coefficient, and B represents the Brownian motion [65]. This approach was applied in early works such as ProteinSGM [41], achieving initial protein backbone generation by modeling full-atom representations, including side chains. Although conventional data types such as images and videos are typically represented in Euclidean space, certain domains like robotics, geosciences, and protein modeling often deal with data inherently defined on Riemannian manifolds [12]. Recognizing this, geometric diffusion models leverage manifold-based representations to capture the intrinsic geometric properties of protein structures more accurately. These models extend the generative process to a manifold M, allowing for a more nuanced approach:

$$dx = [f(x,t) - \frac{1}{2} g(t)^2 \text{grad}_M \log p_t(x)]dt + g(t)dB_M \tag{2}$$

where grad_M and dB_M represent the gradient and Brownian motion on the manifold, respectively [18]. This approach naturally maintains SE(3) invariance and can more effectively represent internal coordinate constraints such as angles and dihedral angles.

In implementation, geometric models typically use geodesic random walks (GRW) for sampling, which requires operating in the tangent space of the manifold and then projecting the results back onto the manifold [35]. The development of manifold diffusion methods has undergone a series of improvements. RGSM systematically extended the score matching to Riemannian manifolds [10]. PNDM proposed pseudo-numerical methods to ensure that the samples remain in the target manifold [43]. RDM generalized continuous-time diffusion models to arbitrary Riemannian manifolds [26].

In addition to manifold-based methods, some researchers have explored graph-based geometric methods such as Graph GDP and NVDiff [27, 15]. These methods use graph structures to represent the geometric relationships of proteins, effectively capturing spatial relationships between atoms or residues. Methods like FrameDiff [86] and RFdiffusion [78] have achieved high-quality protein backbone generation through SE(3) diffusion. These models not only improve the quality of generated structures but also enhance their applicability in various downstream applications. For example, RFdiffusion [78] has demonstrated significant advantages in complex tasks such as motif-scaffolding and binder design through pre-training strategies and conditional generation techniques.

The transition from Euclidean to geometric diffusion models has enhanced protein structure modeling by incorporating intrinsic geometric properties [78]. This paradigm shift, demonstrated through manifold-based and graph-based methods, enables precise structural control while respecting physical constraints.

3.2.2 Transformations of protein structure: equivariance *v.s.* non-equivariant

In protein structures, since the function and properties of protein molecules do not depend on their specific position or orientation in three-dimensional space, it is crucial to ensure that the model is equivariant to geometric transformations such as rotations and translations [58, 8]. This guarantees that the model maintains physical consistency and robustness when dealing with protein structures. Equivariance refers to the property of a model or function where, when the input undergoes a certain transformation, the output changes in a corresponding manner, thereby preserving the fundamental structure and relationships within the data [39, 16]. In the field of molecular modeling, this concept implies that when a molecule rotates or translates, its properties - such as electron density or interatomic forces - should transform accordingly, while maintaining their essential characteristics [5, 60].

From a formal perspective, a function $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$ is deemed equivariant with respect to a set of transformations G (rotations and translations on the original structure) if it satisfies the following condition:

$$\mathcal{F} \circ T_g(x) = S_g \circ \mathcal{F}(x) \tag{3}$$

where T_g and S_g denote transformations corresponding to an element $g \in G$, operating on the vector spaces $\mathcal{X}$ and $\mathcal{Y}$ respectively. Protein structure design focuses on the SE(3) group, which

encompasses rotations and translations in 3D space. This ensures that the model's predictions remain invariant under these transformations, preserving key structural properties [83].

This characteristic has driven significant breakthroughs in diffusion models, enabling more accurate modeling of structural transformations and enhancing the capacity to reconstruct original data with greater fidelity during the reverse diffusion process [24]. The core idea of traditional diffusion models is to gradually add noise to the data through a forward process and then learn a reverse process to recover the data from the noise. The forward process is defined as a Markov chain, where the data at each time step x_t is obtained through a Gaussian distribution $q(x_t|x_{t-1})$, and the reverse process denoises the data using a parameterized reverse transition kernel $p_\theta(x_{t-1}|x_t)$. The key equations of traditional diffusion models are as follows:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \qquad p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I) \tag{4}$$

Anand and Achim's equivariant diffusion model [1] extends the traditional diffusion models framework to address the challenges of rotational and translational symmetry in protein structures. The model introduces Spherical Linear Interpolation (SLERP) to handle rotational variables, allowing the diffusion process to proceed stably in rotationally symmetric spaces. The specific equation is:

$$q_t^i = \text{SLERP}(q_0^i, q_T^i, \alpha_t) \tag{5}$$

The model also integrates Invariant Point Attention (IPA) to maintain rotational and translational invariance and adopts Frame Aligned Point Error (FAPE) as the loss function to accurately address errors in rotational diffusion. These enhancements significantly improve the model's efficiency and precision in generating physically plausible protein structures. Later, researchers have made various improvements to the isovariant diffusion model. In July 2023, Watson *et al.* [78] developed RFdiffusion, demonstrating advantages in motif-scaffolding and binder design through conditional generation techniques. Following this, Bose *et al.* [11] introduced the SE(3)-stochastic flow matching method, improving computational efficiency and generation quality. FrameDiff [84] enhanced computational efficiency through geometric methods, and EigenFold [34] captured global structural features via diffusion in the protein eigenmode space. Equivariant diffusion models have made significant impacts in practical applications. The recently proposed DiffLinker [28], an E(3)-equivariant 3D conditional diffusion model, designs complete molecules from disconnected fragments, enhancing the efficiency and success of fragment-based drug discovery (FBDD) while producing drug-like molecules with high binding affinities.

However, these equivariant models still face challenges in protein structure generation, particularly in maintaining protein chirality. Some non-equivariant models have shown superior performance in addressing this issue. The comparison between equivariant methods and non-equivariant methods is provided in Table 2. While equivariant models preserve rotational and translational symmetry, they often also exhibit reflectional symmetry, which can violate protein chirality, leading to biologically implausible conformations. To solve this problem, Wu *et al.* introduced FoldingDiff [80] in 2024. FoldingDiff [80] achieves translation and rotation invariance by representing protein backbones as sequences of internal angles, thus inherently embedding geometric invariance without relying on complex equivariant architectures. This approach allows a standard Transformer-based diffusion model to iteratively denoise random angles into protein structures, capturing protein chirality and generating diverse, designable structures. FoldingDiff [80] represents a significant advancement in computational protein design by achieving effective modeling without the need for explicit equivariance constraints.

4 Future direction

4.1 Challenges

Validation and evaluation. Despite advances in generative models, validating the quality and functionality of generated protein sequences remains a significant challenge. This difficulty arises in part from the diverse functional requirements of proteins in different biological and industrial contexts.

Table 2: **Comparison of four types of diffusion models in protein design.** The table compares two types of space-based diffusion methods, Euclidean and Geometric, as well as two types of symmetry-based diffusion methods, Equivariant and Non-equivariant, in the context of protein design.

Concept	Method	Description	Advantages	Disadvantages
Euclidean	Anand Achim (2022) [1] ProtDiff (2022) [71] NVDiff (2022) [15]	Operating in Euclidean space using 3D coordinates or distance matrices	High computational efficiency, suitable for rapidly processing large-scale datasets	Struggles to capture complex geometric structures
Geometric	RGSM (2022) [10] Graph GDP (2022) [27] RFdiffusion (2023) [78] FrameDiff (2023) [86] DiffLinker (2024) [28] FoldingDiff (2024) [80]	Leveraging geometric properties like internal angles and local coordinate systems to generate and transform 3D structures	Effectively captures complex geometric structures, enhancing representation of molecular interactions	Computationally intensive, requiring substantial resources for large datasets
Equivariant	Anand Achim (2022) [1] ProtDiff (2022) [71] FrameDiff (2023) [86] RFdiffusion (2023) [78] DiffLinker (2024) [28]	Leveraging SE(3) principles to ensure 3D structures maintain symmetry under rotations and translations	Ideal for tasks needing symmetry and invariance, enhancing reliability in interaction modeling	Higher complexity, requiring advanced mathematical frameworks
Non-equivariant	RGSM (2022) [10] NVDiff (2022) [15] Graph GDP (2022) [27] FoldingDiff (2024) [80]	Using alternative representations to ensure structural validity without equivariant networks	Flexible representation methods enhance structural validity, allowing for diverse modeling approaches	May lack robustness to transformations, potentially affecting performance

For example, proteins optimized for therapeutic use often have distinct requirements compared to those in industrial catalysis, complicating universal definitions. As a result, protein functionality should be evaluated by how well it meets application-specific criteria rather than using a universal benchmark [48]. Ruffolo *et al.* emphasized the need for stronger evaluation criteria to standardize the measurement of protein properties such as stability, solubility, and efficiency in different fields [32]. In addition, wet lab experiments are often necessary to assess basic properties such as stability and expression, as well as design-based properties such as affinity and specificity [92]. This requirement for experimental validation hinders the rapid optimization of the model.

Scarcity of labeled data. Another major challenge in protein modeling is the scarcity of labeled data, particularly for proteins with verified functions. This hinders the development of models that can reliably predict functionality. Acquiring such labeled data is resource intensive and requires significant time, especially when experimental validation is necessary [25, 9, 7]. Many recent studies have highlighted this challenge. Wang *et al.* emphasized the use of generative models to leverage large amounts of unlabeled data alongside smaller labeled datasets [25], while other researchers explored semi-supervised and self-supervised learning techniques that further enhance model performance by minimizing the need for extensive labeled data [9]. In conclusion, while obtaining labeled data remains a bottleneck, utilizing semi-supervised learning methods and relying on the abundance of unlabeled protein sequences offers a promising way to mitigate this challenge.

Complexity and applicability. Current protein generative models are struggling with more complex and applicable tasks. At the intermolecular level, despite the breakthrough in single-chain structure prediction achieved by AlphaFold2, Mardikoraem *et al.* observed that current models still struggle to accurately capture spatial orientations and interactions between chains in multi-chain complexes [36, 50]. Pandy *et al.* further emphasized this challenge in the context of predicting G-protein-coupled receptor complexes [53]. At the intramolecular level, the complexity introduced by post-translational modifications has drawn considerable attention from researchers. For instance, Chai exhibits high sensitivity to protein modifications, with alterations in modified residues often leading to substantial changes in predicted structures [69]. This aligns with the findings of Madani *et al.* and Jing *et al.*,

who highlighted the limitations of current approaches in handling protein diversity [48, 32]. These limitations hinder the design of proteins with precise, tailored functions for advanced applications in medicine, biotechnology, and materials science.

4.2 Opportunities

Leveraging large-scale datasets and pre-trained models. The availability of large datasets, such as the Profluent Atlas with 18 billion protein sequences, offers great opportunities for protein language models (PLM) to improve predictive accuracy and explore de novo protein design. This resource is 10 times larger than AlphaFold's database and enables models to learn complex sequence-structure relationships, thus accelerating advancements in protein design [48, 54]. Such large datasets create an environment where PLMs can identify intricate relationships between sequence and structure, enabling more precise functional predictions. This opportunity is supported by the scaling of generative models in biological research, where PLMs like ProGen2 [47] and others have been shown to generate novel proteins that function just as well as natural proteins optimized by evolution [47].

Integration of advanced techniques. There is significant potential in combining different generative frameworks and methodologies to enhance protein design. Exploring hybrid generative modeling approaches beyond either autoregressive or diffusion models, such as Diffusion Forcing [14], might leads to new paradigm for protein generative modeling. Additionaly, integrating reinforcement learning with PLMs for controllable design shows promise as demonstrated by the success of RL in task-directed tuning of LLMs like ChatGPT [3, 46]. Moreover, implementing a cyclical training process with experimental feedback could be beneficial [21]. This iterative approach underscores the importance of using experimental results to continuously refine model training. Such a strategy could potentially lead to more accurate and functional protein designs, bridging the gap between computational predictions and experimental results.

5 Conclusion

This survey has highlighted key advancements in generative models for protein engineering, focusing on sequence-based, structure-based, and joint sequence-structure generation. Autoregressive models like RNNs, LSTMs, and Transformers have shown promise in capturing complex sequence patterns, while GNNs and diffusion models have excelled in structure generation and co-generation. Despite these successes, challenges remain, including functional validation, scarcity of labeled data, and limitations in handling complex structures like multi-chain complexes and post-translational modifications. Looking forward, we have identified opportunities in leveraging large datasets like the Profluent Atlas and integrating advanced techniques such as reinforcement learning and hybrid modeling. The combination of generative models with experimental feedback has also shown potential to enhance precision in protein design. Continued research in these areas has opened new pathways for breakthroughs in biotechnology, enabling the creation of novel functional proteins.

References

[1] Namrata Anand and Tudor Achim. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *arXiv preprint arXiv:2205.15019*, 2022.

[2] Frances H Arnold. Directed evolution: bringing new chemistry to life. *Angewandte Chemie (International Ed. in English)*, 57(16):4143, 2018.

[3] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[4] Pierre Baldi, Søren Brunak, Paolo Frasconi, Giovanni Soda, and Gianluca Pollastri. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, 15(11):937–946, 1999.

[5] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E (3)-equivariant graph

neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1):2453, 2022.

[6] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

[7] Gokul Bhusal, Ekaterina Merkurjev, and Guo-Wei Wei. Persistent laplacian-enhanced algorithm for scarcely labeled data classification. *arXiv preprint arXiv:2305.16239*, 2023.

[8] Frimpong Boadu, Hongyuan Cao, and Jianlin Cheng. Combining protein sequences and structures with transformers and equivariant graph neural networks to predict protein function. *Bioinformatics*, 39(Supplement_1):i318–i325, 2023.

[9] Frimpong Boadu and Jianlin Cheng. Improving protein function prediction by learning and integrating representations of protein sequences and function labels. *Bioinformatics Advances*, page vbae120, 2024.

[10] Valentin De Bortoli, Emile Mathieu, MJ Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. *Advances in Neural Information Processing Systems*, page 46, 2022.

[11] Avishek Joey Bose, Tara Akhound-Sadegh, Kilian Fatras, Guillaume Huguet, Jarrid Rector-Brooks, Cheng-Hao Liu, Andrei Cristian Nica, Maksym Korablyov, Michael Bronstein, and Alexander Tong. Se (3)-stochastic flow matching for protein backbone generation. *arXiv preprint arXiv:2310.02391*, 2023.

[12] Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z Li. A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[13] Abel Chandra, Laura Tünnermann, Tommy Löfstedt, and Regina Gratz. Transformer-based deep learning for predicting protein properties in the life sciences. *Elife*, 12:e82819, 2023.

[14] Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *arXiv preprint arXiv:2407.01392*, 2024.

[15] Xiaohui Chen, Yukun Li, Aonan Zhang, and Li-ping Liu. Nvdiff: Graph generation through the diffusion of node vectors. *arXiv preprint arXiv:2211.10794*, 2022.

[16] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016.

[17] Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

[18] Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. *Advances in Neural Information Processing Systems*, 35:2406–2422, 2022.

[19] Noelia Ferruz, Steffen Schmidt, and Birte Höcker. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.

[20] Zhangyang Gao, Cheng Tan, and Stan Z Li. Diffsds: A language diffusion model for protein backbone inpainting under geometric conditions and constraints. *arXiv preprint arXiv:2301.09642*, 2023.

[21] Nate Gruver, Samuel Stanton, Nathan Frey, Tim GJ Rudner, Isidro Hotzel, Julien Lafrance-Vanasse, Arvind Rajpal, Kyunghyun Cho, and Andrew G Wilson. Protein design with guided discrete diffusion. *Advances in neural information processing systems*, 36, 2024.

[22] Daniel Hesslow, Niccoló Zanichelli, Pascal Notin, Iacopo Poli, and Debora Marks. Rita: a study on scaling up generative protein sequence models. *arXiv preprint arXiv:2205.05789*, 2022.

[23] S Hochreiter. Long short-term memory. *Neural Computation MIT-Press*, 1997.

[24] Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pages 8867–8887. PMLR, 2022.

[25] Chloe Hsu, Clara Fannjiang, and Jennifer Listgarten. Generative models for protein structures and sequences. *nature biotechnology*, 42(2):196–199, 2024.

[26] Chin-Wei Huang, Milad Aghajohari, Joey Bose, Prakash Panangaden, and Aaron C Courville. Riemannian diffusion models. *Advances in Neural Information Processing Systems*, 35:2750–2761, 2022.

[27] Han Huang, Leilei Sun, Bowen Du, Yanjie Fu, and Weifeng Lv. Graphgdp: Generative diffusion processes for permutation invariant graph generation. In *2022 IEEE International Conference on Data Mining (ICDM)*, pages 201–210. IEEE, 2022.

[28] Ilia Igashov, Hannes Stärk, Clément Vignac, Arne Schneuing, Victor Garcia Satorras, Pascal Frossard, Max Welling, Michael Bronstein, and Bruno Correia. Equivariant 3d-conditional diffusion model for molecular linker design. *Nature Machine Intelligence*, pages 1–11, 2024.

[29] John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. Generative models for graph-based protein design. *Advances in neural information processing systems*, 32, 2019.

[30] Feng Jiao. Probabilistic graphical models and algorithms for. 2008.

[31] Wengong Jin, Jeremy Wohlwend, Regina Barzilay, and Tommi Jaakkola. Iterative refinement graph neural network for antibody sequence-structure co-design. *arXiv preprint arXiv:2110.04624*, 2021.

[32] Bowen Jing, Bonnie Berger, and Tommi Jaakkola. Alphafold meets flow matching for generating protein ensembles. *arXiv preprint arXiv:2402.04845*, 2024.

[33] Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael John Lamarre Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. In *International Conference on Learning Representations*, 2020.

[34] Bowen Jing, Ezra Erives, Peter Pao-Huang, Gabriele Corso, Bonnie Berger, and Tommi Jaakkola. Eigenfold: Generative protein structure prediction with diffusion models. *arXiv preprint arXiv:2304.02198*, 2023.

[35] Erik Jørgensen. The central limit problem for geodesic random walks. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 32(1):1–64, 1975.

[36] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

[37] Nayoung Kim, Minsu Kim, and Jinkyoo Park. Anfinsen goes neural: a graphical model for conditional antibody design. *arXiv preprint arXiv:2402.05982*, 2024.

[38] Suhyeon Kim, Seongmin Ga, Hayeon Bae, Ronald Sluyter, Konstantin Konstantinov, Lok Kumar Shrestha, Yong Ho Kim, Jung Ho Kim, and Katsuhiko Ariga. Multidisciplinary approaches for advanced enzyme biocatalysis in pharmaceuticals: protein engineering, computational biology, and nanoarchitectonics. *EES Catalysis*, 2023.

[39] Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International conference on machine learning*, pages 2747–2755. PMLR, 2018.

[40] Tim Kucera, Matteo Togninalli, and Laetitia Meng-Papaxanthos. Conditional generative modeling for de novo protein design with hierarchical functions. *Bioinformatics*, 38(13):3454–3461, 2022.

[41] Jin Sub Lee, Jisun Kim, and Philip M Kim. Proteinsgm: Score-based generative modeling for de novo protein design. *bioRxiv*, pages 2022–07, 2022.

[42] Matti Leisola and Ossi Turunen. Protein engineering: opportunities and challenges. *Applied Microbiology and Biotechnology*, 75(6):1225–1232, 2007.

[43] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022.

[44] Wenjie Liu, Gengwei Huo, and Peng Chen. Clinical benefit of pembrolizumab in treatment of first line non-small cell lung cancer: a systematic review and meta-analysis of clinical characteristics. *BMC cancer*, 23(1):458, 2023.

[45] Xueliang Liu. Deep recurrent neural network for protein function prediction from sequence. *arXiv preprint arXiv:1701.08318*, 2017.

[46] Ximing Lu, Sean Welleck, Jack Hessel, Liwei Jiang, Lianhui Qin, Peter West, Prithviraj Ammanabrolu, and Yejin Choi. Quark: Controllable text generation with reinforced unlearning. *Advances in neural information processing systems*, 35:27591–27609, 2022.

[47] Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106, 2023.

[48] Ali Madani and Jeffrey A. Ruffolo. Giving structure to language: Profluent's ai models move toward precise and steerable protein design. *GEN News*, 2024.

[49] Mehrsa Mardikoraem, Zirui Wang, Nathaniel Pascual, and Daniel Woldring. Generative models for protein sequence modeling: recent advances and future directions. *Briefings in Bioinformatics*, 24(6):bbad358, 2023.

[50] Nathaniel Pascual Mehrsa Mardikoraem, Zirui Wang and Daniel Woldring. Generative models for protein sequence modeling: recent advances and future directions. *Briefings in Bioinformatics*, 2023.

[51] Geraldene Munsamy, Sebastian Lindner, Philipp Lorenz, and Noelia Ferruz. Zymctrl: a conditional language model for the controllable generation of artificial enzymes. In *NeurIPS Machine Learning in Structural Biology Workshop*, 2022.

[52] Pascal Notin, Ruben Weitzman, Debora Marks, and Yarin Gal. Proteinnpt: Improving protein property prediction and design with non-parametric transformers. *Advances in Neural Information Processing Systems*, 36:33529–33563, 2023.

[53] Gáspár Pándy-Szekeres, Jimmy Caroli, Alibek Mamyrbekov, Ali A Kermani, György M Keserű, Albert J Kooistra, and David E Gloriam. Gpcrdb in 2023: state-specific structure models using alphafold2 and new ligand resources. *Nucleic acids research*, 51(D1):D395–D402, 2023.

[54] Profluent AI. Profluent atlas and large-scale generative protein models, 2024. Accessed: September 2024.

[55] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.

[56] Donatas Repecka, Vykintas Jauniskis, Laurynas Karpus, Elzbieta Rembeza, Irmantas Rokaitis, Jan Zrimec, Simona Poviloniene, Audrius Laurynenas, Sandra Viknander, Wissam Abuajwa, et al. Expanding functional protein sequence spaces using generative adversarial networks. *Nature Machine Intelligence*, 3(4):324–333, 2021.

[57] Domenico Ribatti, Gerardo Cazzato, Roberto Tamma, Tiziana Annese, Giuseppe Ingravallo, and Giorgina Specchia. Immune checkpoint inhibitors targeting pd-1/pd-l1 in the treatment of human lymphomas. *Frontiers in Oncology*, 14, 2024.

[58] Rahmatullah Roche, Bernard Moussad, Md Hossain Shuvo, and Debswapna Bhattacharya. E (3) equivariant graph neural networks for robust and accurate protein-protein interaction site prediction. *PLoS Computational Biology*, 19(8):e1011435, 2023.

[59] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[60] Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning*, pages 9377–9388. PMLR, 2021.

[61] Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306, 2020.

[62] Chence Shi, Chuanrui Wang, Jiarui Lu, Bozitao Zhong, and Jian Tang. Protein sequence and structure co-design with equivariant translation. *arXiv preprint arXiv:2210.08761*, 2022.

[63] Vítor Silva and Cristiano Matos. Recent updates in the therapeutic uses of pembrolizumab: a brief narrative review. *Clinical and Translational Oncology*, pages 1–13, 2024.

[64] Søren Kaae Sønderby and Ole Winther. Protein secondary structure prediction with long short term memory networks. *arXiv preprint arXiv:1412.7828*, 2014.

[65] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[66] Zhenqiao Song, Yunlong Zhao, Wenxian Shi, Yang Yang, and Lei Li. Functional geometry guided protein sequence and backbone structure co-design. *arXiv preprint arXiv:2310.04343*, 2023.

[67] Hannes Stärk, Octavian Ganea, Lagnajit Pattanaik, Regina Barzilay, and Tommi Jaakkola. Equibind: Geometric deep learning for drug binding structure prediction. In *International conference on machine learning*, pages 20503–20521. PMLR, 2022.

[68] Alexey Strokach and Philip M Kim. Deep generative modeling for protein design. *Current opinion in structural biology*, 72:226–236, 2022.

[69] Chai Discovery Team. Chai-1: Technical report. Technical report, Chai Lab, 2024. Technical Report.

[70] John Thomas, Naren Ramakrishnan, and Chris Bailey-Kellogg. Protein design by sampling an undirected graphical model of residue constraints. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(3):506–516, 2008.

[71] Brian L Trippe, Jason Yim, Doug Tischer, David Baker, Tamara Broderick, Regina Barzilay, and Tommi Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. *arXiv preprint arXiv:2206.04119*, 2022.

[72] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[73] Yogesh Verma, Markus Heinonen, and Vikas Garg. Abode: Ab initio antibody design using conjoined odes. In *International Conference on Machine Learning*, pages 35037–35050. PMLR, 2023.

[74] Lara Sellés Vidal, Mark Isalan, John T Heap, and Rodrigo Ledesma-Amaro. A primer to directed evolution: current methodologies and future directions. *RSC Chemical Biology*, 4(4):271–291, 2023.

[75] Jesse Vig, Ali Madani, Lav R. Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. Bertology meets biology: Interpreting attention in protein language models, 2021.

[76] Sheng Wang, Jian Peng, Jianzhu Ma, and Jinbo Xu. Protein secondary structure prediction using deep convolutional neural fields. *Scientific reports*, 6(1):1–11, 2016.

[77] Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *BioRxiv*, pages 2022–12, 2022.

[78] Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

[79] Martin Weigt, Robert A White, Hendrik Szurmant, James A Hoch, and Terence Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009.

[80] Kevin E Wu, Kevin K Yang, Rianne van den Berg, Sarah Alamdari, James Y Zou, Alex X Lu, and Ava P Amini. Protein structure generation via folding diffusion. *Nature communications*, 15(1):1059, 2024.

[81] Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, et al. High-resolution de novo structure prediction from primary sequence. *BioRxiv*, pages 2022–07, 2022.

[82] Zachary Wu, Kevin K Yang, Michael J Liszka, Alycia Lee, Alina Batzilla, David Wernick, David P Weiner, and Frances H Arnold. Signal peptides generated by attention-based neural networks. *ACS Synthetic Biology*, 9(8):2154–2161, 2020.

[83] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923*, 2022.

[84] Jason Yim, Andrew Campbell, Andrew YK Foong, Michael Gastegger, José Jiménez-Luna, Sarah Lewis, Victor Garcia Satorras, Bastiaan S Veeling, Regina Barzilay, Tommi Jaakkola, et al. Fast protein backbone generation with se (3) flow matching. *arXiv preprint arXiv:2310.05297*, 2023.

[85] Jason Yim, Hannes Stärk, Gabriele Corso, Bowen Jing, Regina Barzilay, and Tommi S Jaakkola. Diffusion models in protein structure and docking. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 14(2):e1711, 2024.

[86] Jason Yim, Brian L. Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. Se(3) diffusion model with application to protein backbone generation, 2023.

[87] Qianmu Yuan, Chong Tian, Yidong Song, Peihua Ou, Mingming Zhu, Huiying Zhao, and Yuedong Yang. Gpsfun: geometry-aware protein sequence function predictions with language models. *Nucleic Acids Research*, page gkae381, 2024.

[88] Buzhong Zhang, Jinyan Li, and Qiang Lü. Prediction of 8-state protein secondary structures by a novel deep learning architecture. *BMC bioinformatics*, 19:1–13, 2018.

[89] Cheng Zhang, Adam Leach, Thomas Makkink, Miguel Arbesú, Ibtissem Kadri, Daniel Luo, Liron Mizrahi, Sabrine Krichen, Maren Lang, Andrey Tovchigrechko, et al. Framedipt: Se (3) diffusion model for protein structure inpainting. *bioRxiv*, pages 2023–11, 2023.

[90] Hong Zhang, Jiajing Lan, Huijie Wang, Ruijie Lu, Nanqi Zhang, Xiaobai He, Jun Yang, and Linjie Chen. Alphafold2 in biomedical research: facilitating the development of diagnostic strategies for disease. *Frontiers in Molecular Biosciences*, 11:1414916, 2024.

[91] Pan Zhang, Shuangjia Zheng, Jianwen Chen, Yaoqi Zhou, and Yuedong Yang. Deepanis: Predicting antibody paratope from concatenated cdr sequences by integrating bidirectional long-short-term memory and transformer neural networks. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 118–124. IEEE, 2021.

[92] Ziyi Zhou, Liang Zhang, Yuanxi Yu, Banghao Wu, Mingchen Li, Liang Hong, and Pan Tan. Enhancing efficiency of protein language models with minimal wet-lab data through few-shot learning. *Nature Communications*, 15(1):5566, 2024.