A Survey on Conditional Protein Generation Using Diffusion Models

G.J. Admiraal
TU Delft, 4871669, g.j.admiraal@student.tudelft.nl

ABSTRACT

Proteins, essential to life, exhibit diverse functions shaped by their complex structures of amino acids. Malfunctioning proteins contribute to diseases, while engineered proteins offer potential medical and industrial benefits. However, the vast protein search space poses challenges in design tools and validation. Computational methods, particularly generative models like diffusion models, have emerged as promising tools for protein design. This survey explores how conditional generation within diffusion models facilitates tailored protein properties, guiding the design process towards specific outcomes. Conditioning plays a crucial role in shaping these outcomes, offering potential in targeted therapeutics and industrial applications. The survey covers foundational concepts, conditional settings, recent advancements, and future directions in protein design within diffusion models.

1 INTRODUCTION

Proteins are a ubiquitous and essential tool for any living organism. These intricate biomolecules, comprised of amino acids, fold into unique, complex structures. Proteins take part in numerous biological processes such as catalysing metabolic reactions, aiding the immune system, or adding structure to cells. On the contrary, misfolded or malfunctioning proteins can cause various diseases such as Alzheimer's, Parkinson's, and Huntington's disease.

Proteins are composed of repeating monomer molecules called amino acids. Each amino acid has a standard backbone atom structure of $N - C_\alpha - C$ (see Figure 1). These amino acids vary based on their distinctive side chains, resulting in 20 different types. When hundreds to thousands of amino acids chain together through peptide bonds, they create proteins. The sequence of amino acids dictates the protein's final 3D structure. The structure of a protein, in turn, determines the biological activity and function of the protein.

The design of proteins is driven by their vital role in the functioning of living organisms. This task is motivated by the potential to optimize the functionality of proteins or create new ones. Given the pivotal role of proteins in living organisms, a crucial need for designing proteins arises, whether to enhance their activities or to create new ones. These engineered proteins might unlock opportunities for uncovering innovative methods to leverage cellular pathways. Potentially paving the way for novel treatments aimed at currently untreatable diseases [Koutsopoulos 2017] or for their use in biochemical applications in various industrial settings[Leiman and Taylor 2019]. Unfortunately, the design of proteins faces a major hurdle due to the enormity of the protein search space, which encompasses potential amino acid sequences 20^{100} for a 100 residue protein. Moreover, natural evolution has only explored a limited size of this expansive space [Dryden et al. 2008]. Consequently, there is a broad unexplored design landscape with the potential to reveal entirely new proteins possessing novel properties and functions. However, the sheer size of this design space, coupled with the costs involved in experimental validation [Jaskolski et al. 2014], results in significant challenges in developing effective tools for designing *de novo* protein sequences with specific desired structures and features [Paladino et al. 2017].

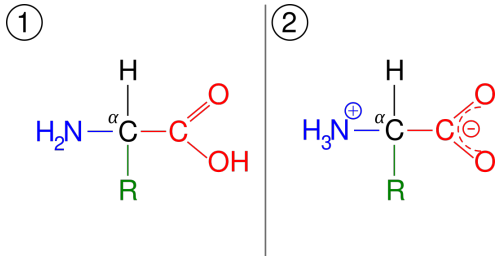

Figure 1: The molecular structure of an amino acid in its (1) un-ionized and (2) zwitterionic forms. Showing the central alpha carbon (black), the carboxyl group (red), the amino group (blue), and the variable side chain (green)[Wikipedia, the free encyclopedia 2012]

Historically, protein design required minimal and rational design approaches whereby the placement of each residue in a design was reasoned using chemical principles and/or biochemical knowledge[Woolfson 2021]. This reliance on labour-intensive experimental techniques required significant expertise. The advent of computational methods and machine learning facilitated a more efficient exploration of the protein search space, enabling the utilisation of these innovative approaches. In recent times, generative models, a subset of deep learning models capable of producing novel outputs following a specified distribution, have captured the attention of protein researchers.

Various generative models have had significant successes in various fields, each offering unique capabilities and applications. Deep Generative Adversarial Networks (GANs), as pioneered by [Goodfellow et al. 2014], involve a dual learning process where two models compete against each other: a generator for crafting novel instances and a discriminator for categorizing them as real or fake. However, one drawback of GANs is that they tend to lack diversity in their output. On the other hand, Variational autoencoders (VAEs), as proposed by [Kingma and Welling 2013], employ an encoder-decoder setup that facilitates easy sampling from the latent space, resulting in more diverse output. Although VAEs can produce more

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

diverse outputs, they often lack quality. Diffusion models, which have gained prominence recently, address these limitations.

Diffusion models were pioneered by [Sohl-Dickstein et al. 2015] [Ho et al. 2020] [Dhariwal and Nichol 2021]. In recent years significant advancements have been made, mainly in the field of computer vision, resulting in state-of-the-art solutions [Nichol et al. 2021] [Rombach et al. 2022] [Ruiz et al. 2023]. These models possess a set of features that is highly relevant for the generation of novel proteins [Watson et al. 2023]. That is, diffusion models exhibit the ability to generate diverse outputs that can be conditionally guided toward specific design objectives, which is not as easily done in other generative models. Furthermore, they possess the capability of inpainting, allowing them to fill in missing portions of partially complete inputs. Lastly, diffusion models offer rotation-equivariant output by employing specialized equivariant architectures.

The generation of proteins involves creating new protein sequences or structures. There are two primary types of protein generation, either by optimising existing proteins or by creating novel proteins. Optimising existing proteins involves refining or modifying existing proteins to improve certain properties, such as binding affinity. The creation of novel proteins requires the design of entirely new proteins. The generation of proteins can involve the generation of desired sequences or structures either partial or complete. Sequence and structure co-design involves generating sequences and structures jointly.

Previous surveys have explored the application of diffusion models in bioinformatics [Guo et al. 2023] and other deep learning methods in protein design [Bennett 2023] [Kim et al. 2023]. A recent survey explored the use of graph diffusion models in molecular, protein, and material design[Zhang et al. 2023]. This survey uniquely focuses on the use of conditional generation within diffusion models for protein design.

Conditioning, within the scope of this survey on protein design, refers to the deliberate introduction of additional protein property information to guide the generative part of the diffusion process toward a specific, intended outcome. In this context, the protein property information significantly shapes the desired result. Rather than randomly selecting a protein from the overall learnt distribution, conditioning ensures that the sample originates from a smaller intended subset within the total distribution. This conditioning information encompasses various protein properties, such as the protein's desired (partial) structures, sequences, or other relevant biophysical attributes. This can also be characteristic information about the (partial) structure or sequences of a binding molecule. Therefore, conditioning is a crucial element of protein design, as it allows designers to tailor protein properties to meet specific objectives.

To illustrate the specific application of this concept, let us consider the following example within a defined context. In protein design, using a structural graph representation of a binding ligand means incorporating specific atom arrangements and bond details crucial for targeting a receptor. For instance, creating a protein for targeted therapy involves ensuring its structure aligns with the depicted ligand graph, enabling precise binding to a specific receptor, like those found on cancer cells, thus enhancing its therapeutic potential.

This survey begins by providing background on the foundational concepts of diffusion models and how they are used in the context of protein design in section 2. Subsequently, we explore various conditional settings in the sections 3, 4, 5 and 6. Section 7 highlights recent advancements in related fields and other future directions for protein design. Finally, we conclude this survey in section 8.

2 BACKGROUND

2.1 Diffusion models

Diffusion models try to learn a data distribution by slowly adding noise to its input and then trying to systematically remove that noise. These models can utilise various architectures tailored to their specific requirements, adapting the approach of noise addition and removal for optimal performance. By understanding the process of removing the noise, the model can generate novel outputs of the data distribution.

Three subtypes exist within this category: Denoising Diffusion Probabilistic Models (DDPMs) [Sohl-Dickstein et al. 2015], Score-based Generative Models (SGMs) [Song and Ermon 2019], and Stochastic Differential Equations (SDEs) [Song et al. 2020]. These subtypes vary in their approaches to executing both the forward and backward diffusion passes. In this section, we will only discuss the DDPMs and SDEs since only these models are used in the research discussed in this paper. An overview of DDPMs and SDEs can be found in figure 2.

2.1.1 Denoising Diffusion Probabilistic Models(DDPM). A Denoising Diffusion Probabilistic Model (DDPM) is a type of generative model capable of creating new *discrete* data samples from a specified data distribution, using a dual Markov chain approach.

In the DDPM framework, the first stage involves the forward diffusion process, which iteratively transforms the original distribution across a specified number of steps, denoted as T. This transformation gradually introduces noise, ultimately converging toward a simpler prior distribution, often a Gaussian distribution. The reason for transforming to such a distribution is that this distribution can be used for easy sampling later. Notably, the amount of noise added at each step is controlled by a predefined variance schedule denoted as β. Formally, the forward process is defined by the posterior probability $q(x_t|x_{t-1})$, where x_t signifies the original input with noise corresponding to the time step t. The full forward process can be defined as follows:

$$q\left(x_{1:T} \mid x_0\right) = \prod_{t=1}^{T} q\left(x_t \mid x_{t-1}\right), \tag{1}$$

$$q\left(x_t \mid x_{t-1}\right) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right) \tag{2}$$

Where $\beta_t \in [0, 1]$ is linked to the variance schedule. Using $\alpha_t = 1 - \beta_t$ and $\bar{\alpha_t} = \prod_{s=1}^{t} \alpha_s$ we can rewrite the previous equation to:

$$q(x_t|x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}x_0, \left(1-\bar{\alpha}_t\right) I\right) \tag{3}$$

Lastly, the backward diffusion process uses a neural network architecture θ that learns to predict the noise that was added in a forward step. This backward process is designed to reconstruct

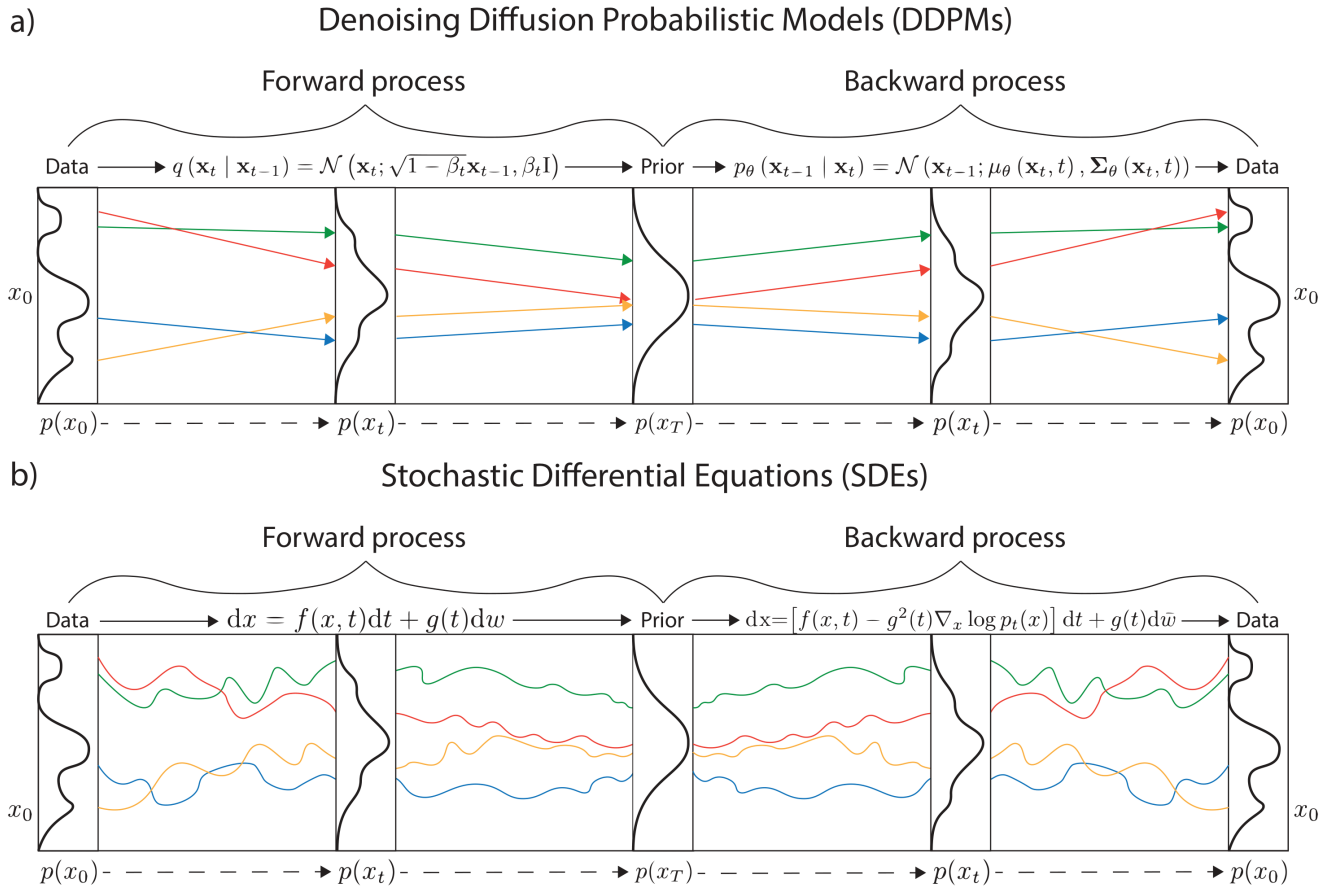

Figure 2: Overview of DDPMs (a) and SDEs (b). A forward process can map our complex data distribution to a noise distribution (the prior). The backward process reverses this noising for generative modelling. During these processes use DDPMs (a) discrete steps and SDEs continuous steps.

the original input based on the predicted noise at each time step. The backward process is formally given as $p_\theta(x_{t-1}|x_t)$, and the optimisation of the model is guided by the following objective.

$$p(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{4}$$

Where $\mu_\theta(x_t, t)$ and $\Sigma_\theta(x_t, t)$ predict the mean and the variance of the noise at time t respectively. In practice, the variance is often kept fixed and only the mean is predicted. The works by [Ho et al. 2020] gave us a simplified version of the objective denoted below:

$$L_{\text{simple}}(\theta) := \mathbb{E}_{t,\mathbf{x}_0,\epsilon}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t\right)\right\|^2\right] \tag{5}$$

2.1.2 Stochastic Differential Equations (SDEs). Stochastic Differential Equations (SDEs) are a class of mathematical models that illustrate how a system evolves amidst random noise. In the research carried out by [Song et al. 2020], the authors revealed the connection between SDE and SGM along with DDPM which lies in their fundamental principles. SDEs, as a framework, support the dynamics of probabilistic modelling in both SGMs and DDPMs, allowing the generation of data by specifying the evolution of probability distributions over time via stochastic processes, forming the basis for these generative models. While SDEs naturally handle continuous data due to their continuous-time nature, DDPMs discretise these continuous-time models, adapting the framework to generate discrete data sequences.

A Score-based Stochastic Differential Equation (Score SDE) is a type of SDE where the drift term is defined as the negative gradient of a score function, and the diffusion term is a function of time as defined by [Song et al. 2020]. A score function $\nabla_x \log(p(x))$ represents the gradient of the log probability density function to the data x.

A forward Stochastic Differential Equation (SDE) is a mathematical framework that shows how a variable changes over time, considering both predictable and random factors in a continuous-time setting. It is characterised as follows:

$$dx = f(x,t)dt + g(t)dw \tag{6}$$

A reverse Stochastic Differential Equation (SDE) describes a continuous-time system that operates in a backward manner, often used to compute the score function for the forward SDE. This score function is utilised to generate samples from the conditional distribution. The reverse SDE is typically defined as follows.

$$\mathrm{d}x = \left[f(x,t) - g^2(t)\nabla_x \log p_t(x)\right]\mathrm{d}t + g(t)\mathrm{d}\bar{w} \tag{7}$$

The score function is approximated by parameterisation in a score model, denoted as $s_\theta(xt, t)$. This process extends the goal of scoring matching to continuous time, as specified by [Yang et al. 2023].

$$\mathbb{E}_{t\sim\mathcal{U}[0,T]),\mathbf{x}_t\sim q(\mathbf{x}_t|\mathbf{x}_0)}\left[\lambda(t)\left\|s_\theta\left(\mathbf{x}_t,t\right) - \nabla_{x_t}\log q_{0t}\left(\mathbf{x}_t \mid \mathbf{x}_0\right)\right\|^2\right] \tag{8}$$

2.1.3 Incorporating Conditional Information. To generate samples that exhibit specific characteristics or conform to particular conditions, these models often incorporate conditional information to influence the generation process. Importantly, the forward diffusion process only adds noise to data so it is irrelevant to data or contexts but the generative diffusion process depends on the given condition and full observation of the previous step. This conditioning can be applied in various ways, allowing the generation of samples to adhere to specific criteria or constraints.

One method involves utilizing conditional distributions, such as $p(x_t \mid y)$, where x_t represents the sample at time step t, and y denotes the conditioning information. This approach enables the model to generate samples based on given conditional data, like class labels, attributes, or any other relevant information. This conditional distribution can be modelled using a separate classifier, heuristic or some approximation.

Furthermore, these models can be guided or conditioned by using a wide array of input data, ranging from simple labels or attributes to more complex structured data. This guidance assists the model in learning and generating samples that align with the provided conditions, making the generated outputs more controllable and tailored to specific requirements. These input data are often embedded and applied using some attention-based module [Vaswani et al. 2017].

In summary, DDPM and SDE are distinct generative modelling techniques with different underlying principles and applications. DDPM focuses on auto-regressive modelling and discrete data, while SDE models the continuous dynamics of data and is more versatile in terms of data types. The choice between them depends on the specific problem and the nature of the data with which one is working.

2.2 Datasets, Post-Processing and Validation

Dataset selection, post-processing and validation of proteins play a crucial role in the accurate generation and analysis of conditional protein structures using diffusion models.

2.2.1 Protein Data Bank. The Protein Data Bank [Berman et al. 2000] serves as a primary source of protein structures. Several specific subsets of this database can be used for their design-specific tasks, such as SAbDab encompasses antibodies or CATH which provides a hierarchical classification of protein domains based on their folding patterns. Since a diffusion model can use different architectures, the used data must match the selected architecture and vice versa.

2.2.2 Post-Processing. The generated proteins often are not specified as full-atom proteins. Several techniques can be utilized to refine and complete the generated protein fully. Protein structure refinement and enhancement can be done using post-processing methods such as ADMM (Alternating Direction Method of Multipliers) [Boyd et al. 2011] and Rosetta [DiMaio et al. 2009]. Additionally, side chain packing algorithms [Alford et al. 2017] can be used to determine the correct placement of side chains and obtain reliable and realistic protein conformations.

2.2.3 Validation. Various techniques validate generated proteins. In silico validation involves a range of computational assessments, such as sequence novelty analysis using BLAST[Altschul et al. 1990], energy calculations, structural analysis using Root-Mean-Square Deviation (RSMD), and advanced structure prediction techniques like AlphaFold [Jumper et al. 2021]. These methods collectively assess and confirm the stability, reliability, and accuracy of the protein structures generated. Additionally, in vitro validation involves experimental methods such as biophysical techniques in a lab (e.g., X-ray crystallography, NMR spectroscopy, cryo-EM) to experimentally validate protein structures, along with functional evaluation to confirm protein function and properties.

2.3 Nomenclature

A table comprising common abbreviations and acronyms relevant to conditional protein design via diffusion models can be seen in table 2. This table serves as a quick reference guide, offering an inventory of frequently used terms and acronyms within this article.

Table 1: Nomenclature table

DDPM	Denoising Diffusion Probabilistic Model
SDE	Stochastic Differential Equation
SGM	Score-based Generative Models
HDM	Harmonic Diffusion Model [Jing et al. 2023]
EGNN	Equivariant Graph Neural Network
ETNN	Equivariant Transformer Neural Network
MLP	Multi-Layer Perceptron
ECNN	Equivariant Convolutional Neural Network
PLM	Protein Language Model
APMixer	Aligned Protein Mixer [Martinkus et al. 2023]
RMSD	Root-Mean-Square Deviation
CDR	Complementarity-determining Regions
BLAST	Basic Local Alignment Search Tool

3 PROTEIN DESIGN THROUGH CONDITIONING ON ITS STRUCTURE PROPERTIES

The structure of a protein is classified into four levels: the primary structure, which denotes the amino acid sequence without 3D considerations; the secondary structure, defining localised folding patterns or motifs such as alpha helices and beta-pleated sheets over several dozen amino acids; the tertiary structure, describing the intricately folded state of the entire amino acid chain; and, in

Table 2: Tabulated summary of reviewed papers. The initial column displays the contributing authors. The next column highlights the Diffusion types, see table 1 for context. The third column specifies the architecture utilised in the backward diffusion process; see table 1 for context. The column of generative type distinguishes between de novo protein generation, optimisation of existing proteins, or generation of dynamic structures. Following that, the table indicates the various conditioning types used, including structure (CoStr), sequence (CoSeq), interactive binding molecule (CoBM) or protein family (CoPF). The final column denotes whether the model primarily predicts sequences, predicts structures or applies a co-design approach.

Authors	Diffusion Type	Architecture	Generative type	Conditioning type	Generates	Code available
[Ni et al. 2023]	DDPM	UNET	De novo	CoStr	Sequence	yes
[Trippe et al. 2022]	DDPM	EGNN	De novo and optimization	CoStr	Structure	yes
[Anand and Achim 2022]	DDPM	ETNN	De novo and optimization	CoStr	Structure	no
[Jing et al. 2023]	HDM	EGNN	Dynamic structure generation	CoSeq	Structure	yes
[Qiao et al. 2022]	SDE	EGNN	Dynamic structure generation	CoSeq, CoBM	Structure	no
[Nakata et al. 2023]	DDPM	EGNN	Dynamic structure generation	CoSeq, CoBM	Structure	yes
[Martinkus et al. 2023]	DDPM	APMixer	De novo	CoPF	Co-design	yes
[Luo et al. 2022]	DDPM	MLP	De novo and optimisation	CoBM	Co-design	no
[Ketata et al. 2023]	SDE	ECNN	Dynamic structure generation	CoBM	Structure	yes
[Watson et al. 2023]	DDPM	RoseTTAFold	De novo and optimisation	CoStr, CoBM	Structure	yes

cases where proteins consist of multiple amino acid sequences, the quaternary structure, determined by the arrangement of these chains.

Conditioning protein design on motifs or secondary structures offers several advantages. One benefit is the influence of motifs on the mechanical properties of protein materials, making it advantageous to use specific motifs for the design of proteins with desired mechanical characteristics. Another advantage lies in the flexibility of this approach, as it avoids excessively restricting the model to generate only one structure, allowing considerable variation while adhering to specified secondary structure constraints.

The following subsections will explore the use of diffusion models to generate protein designs conditioned on various structural information. A brief overview can be found in table 2.

3.1 Novel protein generation conditioned on secondary structure information

[Anand and Achim 2022] introduced an equivariant DDPM for generating protein structure backbones based on topological structure constraints embedded and incorporated using Invariant Point Attention (IPA) [Jumper et al. 2021] modules. On top of these generated backbones, they can diffuse the protein's sequence and rotamers, the orientation of a side chain. This allows them to generate full-atom proteins which do not need a post-processing step, making their model completely end-to-end. Their obtained results show that their model can successfully do region recovery or optimization using inpainting. Furthermore, the model also has comparable results when generating sequences and rotamers with baseline models. Since their model can only generate these sequences and rotamers from previously generated backbones, the model is not able to co-design a complete protein, impacting the model's qualitative results by constraining structure, sequence, and rotamers separately

The work by [Ni et al. 2023] utilizes a DDPM to create new protein sequences based on secondary structure data, bypassing the atomic backbone construction to concentrate on the correlation between secondary structures and sequences. Their model generates sequences by conditioning on a fractional distribution over 8 different types of secondary structures that are embedded and used using cross-attention. It can also take in more specific conditional information in the form of per-residue secondary-structure information, which is encoded and then concatenated to the input. Post-generation, folding prediction methods are employed to forecast complete protein structures and classify their secondary structures, showing alignment with the conditioning information. Lastly, their generated proteins are validated on novelty via BLAST analysis, which highlights the effectiveness of using the per-residue information in generating de novo sequences with $50\% - 60\%$ similarity. Notable, from their work is that there is no mention of equivariant modelling which is a necessary aspect when designing proteins.

A novel multi-purpose model, called RFDiffusion, was recently proposed by [Watson et al. 2023]. Their model harnesses the success of the structure prediction models by fine-tuning it on protein structure denoising tasks. Specifically, the authors used RosettaFold [Baek et al. 2021], but in theory, any of such structure prediction models could work. The model represents C_α coordinates and $N - C_\alpha - C$ rigid orientation of each residue using the RFframe representation. RFDiffusion is capable of conditioning its protein generation on several tasks. RFDiffusion stands out for its ability to generate symmetric oligomers conditioned on specified point group symmetries. These structure symmetry specifications are used as a heuristic during the backward diffusion process. RFDiffusion exhibits high success rates in both in silico predictions and experimental validation. Developing a novel approach to protein structures has shown superior performance in designing diverse protein structures across various symmetries, including those not typically observed in protein structure databases.

3.2 Protein scaffold structure generation conditioned on motif structure

In protein design, scaffolds serve as a stable framework that supports the structural integrity of a specific motif, where motifs are functional protein fragments that contribute to biological functions within the stable structure of the scaffold.

To this end, [Trippe et al. 2022] proposed a two-fold solution. First, they introduced ProtDiff, an unconditional equivariant diffusion model for protein structure sampling. Second, they developed

SMCDiff, which uses sequential Monte Carlo (SMC) sampling in tandem with ProtDiff to conditionally sample scaffolds based on a given motif. During the denoising stage, the authors used an approximation method to estimate the conditional probability of a scaffold given a motif. From their limited results, they could conclude that their two-fold solution is the first one capable of generating diverse scaffolds of more than 20 residues with a significantly lower computation time than other previous methods. Notably, they acknowledge the limitation of evaluating their model based on training data, recognizing the absence of a standard evaluation benchmark for the motif-scaffold protein problem.

The RFDiffusion model [Watson et al. 2023] later reached outstanding results on various motif scaffolding tasks. During training, the conditional probability of a scaffold given a motif is learnt and used during denoising. For functional-motif scaffolding, they proposed a benchmark on which RFDiffusion had an almost perfect score and 21% higher score than the next best model. When their model was tested on scaffolding enzyme active sites, it showed that it was able to scaffold enzyme active sites with high success rates across a range of enzyme classes. Lastly, the model's capabilities were tested on symmetric functional-motif scaffolding, where the scaffold is symmetric, for which it showed in its four test cases that it could design these symmetric scaffolds.

4 PROTEIN DESIGN CONDITIONED ON INTERACTIVE MOLECULAR PARTNERS

Designing proteins with a focus on their interactions with various molecules, such as ligands, antigens, and other proteins, is of significance in the fields of molecular biology and biotechnology. Protein functionality often relies on specific interactions with other molecules. Understanding and harnessing these interactions is crucial for the development of novel therapeutics and biotechnological applications. An example of how the structure of a protein can change when bound to a ligand can be seen in figure 3. A compact list of the papers discussed can be found in table 2.

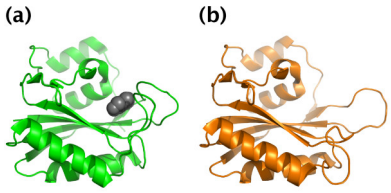

Figure 3: Example of a protein-ligand structure pair between ligand-bound (a) and unbound (b) states [Morita et al. 2011]

4.1 Protein binding-state structure generation conditioned on ligand information

In a diffusion setting, the works by [Qiao et al. 2022] were the first to cast the problem of finding a protein structure given a ligand structure. A ligand is a molecule, often a small chemical compound, that binds to the active site of a protein, modulating the protein's function or activity. NeuralPlexer uses attention modules together with the conditional information to aid the denoising stage. This information includes the protein sequence, structural information obtained from a protein language model (PML), and ligand graph representation. The NeuralPlexer method, when applied to binding site structure recovery, showcased a notable success rate in predicting accurate binding pocket structures within a defined radius around the ligand atom coordinates.

The research of [Nakata et al. 2023] modelled the protein-ligand interaction using an EGNN to generate protein structures conditioned on the protein sequence and the graph representation of the target ligand. During each denoising step, conditional information is embedded and appended to the noise input. Additionally, they generate proteins without protein structural information, unlike NeuralPlexer, and from their results they showcased that this still results in properly generated structures. They then compared their model against other molecular docking methods and showed that their model had comparable or better results than these models. The model seems especially effective compared to other models when dealing with ligands of larger size.

The authors from the RFDiffusion model [Watson et al. 2023] fine-tuned their model to design binders to target molecules, both with and without conditioning on compatible fold information. Conditioning information is used during denoising as the gradient of a heuristic rather than using a trained classifier. The greatest contribution of their results is that high-affinity binders can be identified from dozens of designs instead of several thousand for previous models.

Other notable research, such as DiffDock [Corso et al. 2022] and DiffBP [Lin et al. 2022], has also employed diffusion-based techniques to explore protein-ligand interactions. However, it is essential to note that these methods primarily concentrate on generating potential ligand poses for given ligand-protein pairs and do not directly address protein design. Consequently, they fall outside the scope of this study.

4.2 Conditionally generating protein-protein interaction structures

[Ketata et al. 2023] were inspired by the DiffDock model and focused on rigid protein-protein docking using the DiffDock model. In their work, they generate binding protein poses while keeping the receptor protein fixed. Notable is that they only consider protein structures in their rigid bounded state and keep internal bonds, angles and torsion angles fixed during generation. When only generating one sample their model outperforms the majority of the baseline models. Since its model is generative of nature, it generates a distribution of possible protein-protein complexes. When they select the complex with the smallest RMSD from their generated samples, DiffDock-PP performance exceeds that of all baseline models by a large margin. The authors mention that due to time constraints, they were not able to evaluate DiffDock-PP on the Docking Benchmark 5.5 (DB5.5) dataset, which could lead to different model results. Additionally, from their paper, it is unclear how the conditioning information is provided to the diffusion model.

One other form of protein-protein interaction is the interaction between antibodies and antigens. Antibodies are specialised proteins created by the immune system and designed to bind to specific foreign entities called antigens. These proteins contain specialized regions called complementarity-determining regions (CDRs) that bind to particular parts of antigens, allowing the immune system to

identify, neutralize, or mark them for removal. The strategic design of CDRs is necessary for crafting medical antibodies tailored to target known antigens.

One diffusion-based approach that stands out in the design of these CDRs is [Luo et al. 2022]. Their model, called DiffAb, is explicitly conditional on the protein complex, consisting of an antigen and an antibody framework, allowing it to be generalised to new antigens. The model jointly diffuses the three components of a protein, namely the amino acid types, the C_α coordinates and the amino acid orientations, allowing for co-design of the full protein. DiffAb directly learns the conditional distributions, since conditioning is included in the objective function. The model is tested in CDR inpainting on three antibodies and compared with other models. When testing the model's ability to recover and co-design recover CDRs, it has competitive results. It also shows that their model has reasonably good binding energy on the Rosetta validation method without explicitly being learned on this validation method, unlike other models. Additionally, next to CDR recovery DiffAb is also able to optimize an antibody binding energy. Lastly, being a diffusion model DiffAb shows that it can generate more diverse structures, suggesting a broader structural exploration capability.

5 CONDITIONALLY GENERATING SPECIFIC PROTEIN FAMILIES

Motivated by the observation that key large protein families, typically have strong properties, such as being able to be mapped to a reliable sequence ordinate via sequence alignment, AbDiffuser was developed [Martinkus et al. 2023]. They do this by incorporating family-specific priors during the diffusion process. Their solution incorporates priors from the antibody family and consequently generates antibodies. Their approach differs from existing antibody design methods since AbDiffuser extends beyond CDR generation to encompass the full antibody structure. Claiming that generating complete antibodies broadens design possibilities, such as optimising stability or immunogenicity, and potentially impacting antigen interaction and CDR conformation. The results show that AbDiffuser demonstrates robust antibody generation with lower memory requirements than previous models, despite having more model parameters. Furthermore, from their in-vitro wet lab experiments, they can conclude that their generated antibodies have a higher binding affinity than previous models while using significantly fewer samples. Lastly, they state that their model can also generalize to other large protein families but this warrants further research. A high-level overview of AbDiffuser can be found in table 2.

6 CONDITIONAL PROTEIN DYNAMIC STRUCTURE GENERATION

Protein structure prediction has witnessed remarkable advancements since AlphaFold2 achieved experimental-level accuracy [Jumper et al. 2021]. Following its success subsequent models such as RoseTTAFold [Baek et al. 2021], ESMFold [Lin et al. 2023], HelixFold-Single [Fang et al. 2023] and OmegaFold [Wu et al. 2022] have either replicated or approached similar levels of performance. These newer models make use of protein language model (PLM) representations to extract feature data from the protein sequence to predict the protein structure.

Although these methods excel at modelling static experimental structures derived from crystallography or cryo-electron microscopy (Cryo-EM) data, proteins in their natural environments are not static and exhibit dynamic structural ensembles. This dynamic behaviour is caused by interactions with other molecules and causes the protein's structure and function to change. These changes can initiate various important reactions that help control biological functions.

The following subsections will explore the use of diffusion models for conditional protein structure generation. A compact list of the papers discussed can be found in table 2.

6.1 Protein structure distribution generation conditioned on sequence

To this end, various research has looked at the distributional modelling properties of diffusion models to find the dynamic structure of a protein given its sequence.

The novel model, NeuralPlexer [Qiao et al. 2022], was first proposed to generate protein structure distributions. Specifically, it was tested on ligand-binding proteins that exhibit large conformational variability and the results show that NeuralPlexer gives the highest performance compared to the best-performing structure prediction methods. These other best-performing methods do not take into account crucial ligand information. This shows from their results that this additional conditional ligand information is necessary when predicting the structure of ligand-binding proteins.

Another diffusion-based model that generates dynamic protein-ligand structures conditioned on the sequence was developed by [Nakata et al. 2023]. The outcomes produced by their model showcased a variety of structures, with accurate protein conformations and binding positions for ligands. Although giving good results, this model has been limited by producing proteins of limited size and a lack of training data.

Lastly, there is Eigenfold [Jing et al. 2023], which uses a novel diffusion process called harmonic diffusion and eigenmodes to generate dynamic protein structures from a fixed protein sequence. From their research, it is still unclear how the sequence information is incorporated into their method. When testing their model for single-structure prediction, it is only comparable to RoseTTAFold [Baek et al. 2021] but still inferior to AlphaFold2[Jumper et al. 2021] and ESMFold [Lin et al. 2023]. The results from tests on conformational diversity show that Eigenfold is lacking as well, being unable to generate highly accurate protein models that effectively represent the full range and specific details of protein conformational changes, thereby showing limitations in accurately capturing the diverse structural variations.

7 FUTURE WORK

Many works in the field of protein design often encounter a lack of comprehensive training data, which emphasizes the importance of the generation of viable artificial protein generation. Most current papers tend to train their models on a relatively limited number of proteins, and testing is frequently confined to only a handful of proteins. This scarcity of available viable proteins impedes the development of well-rounded models.

One promising direction for advancing protein design involves establishing a standardized methodology for testing and evaluating protein designs. Given the varied ways of validating proteins, authors might selectively choose metrics that favour their models. To address this, it would be beneficial to propose benchmarks tailored to specific design categories. This approach would enable methods to be compared on a level playing field, encouraging healthy competition among researchers and leading to new advancements in the field of protein design, similar to the impact of standardized benchmarks in various other domains of deep learning [Richter et al. 2017].

Regarding the representation of proteins, the field could benefit from standardization in the way proteins are represented. Presently, different approaches utilize varying neural network architectures that require diverse representations. Standardizing this process would enhance the consistency and comparability of different methodologies.

Additionally, most methods discussed in this paper focus solely on generating either protein sequences or structures. For example, a crucial aspect of enzyme design involves identifying the optimal arrangement of side chains around the binding site to stabilize the transition state. Exploring methods that integrate co-design principles and adopt an end-to-end approach, addressing both sequence and structural aspects simultaneously, would produce better proteins and further advance the field of protein design.

One aspect that the current discussed literature lacks is explicit conditioning on biochemical and biophysical properties, such as binding affinity or temperature degradation. These characteristics are useful when utilising proteins in a controlled industrial setting. Optimizing certain properties of proteins has cost-saving benefits in these settings.

Since diffusion models are commonly utilized and facilitate the most progress in computer vision it could be of value to take inspiration from the progress in this field [Croitoru et al. 2023]. One example is ControlNet [Zhang and Agrawala 2023], a neural network architecture that adds spatial conditioning controls to large, pre-trained text-to-image diffusion models. This allows the model to use edge maps which need to be in the final output of the model. This spatial conditioning could be beneficial for the structural generation of proteins. Other notable research that could be beneficial has focused on faster sampling techniques [Zhang and Chen 2022] or on different conditional guidance methods [Singh et al. 2022]

Future research in conditional diffusion models for protein design can leverage success from the success these models have had in the field of molecular design. One example is EDM (Energy-Based Diffusion Model) [Hoogeboom et al. 2022], which introduces diverse noise for sampling molecular conformations, serving as a foundation for accurate structure generation. Furthermore, DiffSBDD (Structure-Based Drug Design) [Schneuing et al. 2022] employs equivariant DDPM to generate small molecule ligands with high specificity to protein pockets, presenting a powerful tool for crafting high-quality ligands which could be used in collaboration with the design of ligand-binding proteins. Lastly, traditionally diffusion methods have used isotropic Gaussian noise for the forward process. Applications for molecule structure generation have increasingly featured non-isotropic or non-Euclidean processes that exploit the reduced degrees of freedom and chemical priors in a molecular structure, which could warrant further research for protein design as well.

Lastly, since both molecular design and protein design model the correct placement of atoms one future direction could be a multi-purpose model which could do both. Such a general-purpose method would streamline the design process by integrating the principles of molecular and protein design, offering a comprehensive tool capable of accommodating diverse atomic structures.

8 CONCLUSION

conclusion

REFERENCES

Rebecca F Alford, Andrew Leaver-Fay, Jeliazko R Jeliazkov, Matthew J O'Meara, Frank P DiMaio, Hahnbeom Park, Maxim V Shapovalov, P Douglas Renfrew, Vikram K Mulligan, Kalli Kappel, et al. 2017. The Rosetta all-atom energy function for macromolecular modeling and design. *Journal of chemical theory and computation* 13, 6 (2017), 3031–3048.

Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. 1990. Basic local alignment search tool. *Journal of molecular biology* 215, 3 (1990), 403–410.

Namrata Anand and Tudor Achim. 2022. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *arXiv preprint arXiv:2205.15019* (2022).

Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. 2021. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373, 6557 (2021), 871–876.

Nathaniel Richard Bennett. 2023. *Deep Learning Tools for Protein Binder Design.* Ph.D. Dissertation. University of Washington.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. 2000. The protein data bank. *Nucleic acids research* 28, 1 (2000), 235–242.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning* 3, 1 (2011), 1–122.

Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. 2022. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776* (2022).

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* 34 (2021), 8780–8794.

Frank DiMaio, Michael D Tyka, Matthew L Baker, Wah Chiu, and David Baker. 2009. Refinement of protein structures into low-resolution density maps using rosetta. *Journal of molecular biology* 392, 1 (2009), 181–190.

David TF Dryden, Andrew R Thomson, and John H White. 2008. How much of protein sequence space has been explored by life on Earth? *Journal of The Royal Society Interface* 5, 25 (2008), 953–956.

Xiaomin Fang, Fan Wang, Lihang Liu, Jingzhou He, Dayong Lin, Yingfei Xiang, Xiaonan Zhang, Hua Wu, Hui Li, and Le Song. 2023. HelixFold-Single: MSA-free Protein Structure Prediction by Using Protein Language Model as an Alternative. (2023). arXiv:q-bio.BM/2207.13921

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

Zhiye Guo, Jian Liu, Yanli Wang, Mengrui Chen, Duolin Wang, Dong Xu, and Jianlin Cheng. 2023. Diffusion Models in Bioinformatics: A New Wave of Deep Learning Revolution in Action. *arXiv preprint arXiv:2302.10907* (2023).

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

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

Mariusz Jaskolski, Zbigniew Dauter, and Alexander Wlodawer. 2014. A brief history of macromolecular crystallography, illustrated by a family tree and its N obel fruits. *The FEBS journal* 281, 18 (2014), 3985–4009.

Bowen Jing, Ezra Erives, Peter Pao-Huang, Gabriele Corso, Bonnie Berger, and Tommi Jaakkola. 2023. EigenFold: Generative Protein Structure Prediction with Diffusion Models. *arXiv preprint arXiv:2304.02198* (2023).

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko,

et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 7873 (2021), 583–589.

Mohamed Amine Ketata, Cedrik Laue, Ruslan Mammadov, Hannes Stärk, Menghua Wu, Gabriele Corso, Céline Marquet, Regina Barzilay, and Tommi S Jaakkola. 2023. DiffDock-PP: Rigid Protein-Protein Docking with Diffusion Models. *arXiv preprint arXiv:2304.03889* (2023).

Jisun Kim, Matthew McFee, Qiao Fang, Osama Abdin, and Philip M Kim. 2023. Computational and artificial intelligence-based methods for antibody development. *Trends in Pharmacological Sciences* (2023).

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

Sotirios Koutsopoulos. 2017. *Peptide applications in biomedicine, biotechnology and bioengineering*. Woodhead Publishing.

Petr G Leiman and NM Taylor. 2019. Reference Module in Life Sciences. (2019).

Haitao Lin, Yufei Huang, Meng Liu, Xuanjing Li, Shuiwang Ji, and Stan Z Li. 2022. Diffbp: Generative diffusion of 3d molecules for target protein binding. *arXiv preprint arXiv:2211.11214* (2022).

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379, 6637 (2023), 1123–1130.

Shitong Luo, Yufeng Su, Xingang Peng, Sheng Wang, Jian Peng, and Jianzhu Ma. 2022. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. *Advances in Neural Information Processing Systems* 35 (2022), 9754–9767.

Karolis Martinkus, Jan Ludwiczak, Kyunghyun Cho, Wei-Ching Lian, Julien Lafrance-Vanasse, Isidro Hotzel, Arvind Rajpal, Yan Wu, Richard Bonneau, Vladimir Gligorijevic, et al. 2023. AbDiffuser: Full-Atom Generation of In-Vitro Functioning Antibodies. *arXiv preprint arXiv:2308.05027* (2023).

Mizuki Morita, Tohru Terada, Shugo Nakamura, and Kentaro Shimizu. 2011. BUDDY-system: A web site for constructing a dataset of protein pairs between ligand-bound and unbound states. *BMC Research Notes* 4 (2011), 1–4.

Shuya Nakata, Yoshiharu Mori, and Shigenori Tanaka. 2023. End-to-end protein–ligand complex structure generation with diffusion-based generative models. *BMC bioinformatics* 24, 1 (2023), 1–18.

Bo Ni, David L Kaplan, and Markus J Buehler. 2023. Generative design of de novo proteins based on secondary-structure constraints using an attention-based diffusion model. *Chem* (2023).

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).

Antonella Paladino, Filippo Marchetti, Silvia Rinaldi, and Giorgio Colombo. 2017. Protein design: from computer models to artificial intelligence. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 7, 5 (2017), e1318.

Zhuoran Qiao, Weili Nie, Arash Vahdat, Thomas F Miller III, and Anima Anandkumar. 2022. Dynamic-backbone protein-ligand structure prediction with multiscale generative diffusion models. *arXiv preprint arXiv:2209.15171* (2022).

Stephan R Richter, Zeeshan Hayder, and Vladlen Koltun. 2017. Playing for benchmarks. In *Proceedings of the IEEE International Conference on Computer Vision*. 2213–2222.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22500–22510.

Arne Schneuing, Yuanqi Du, Charles Harris, Arian Jamasb, Ilia Igashov, Weitao Du, Tom Blundell, Pietro Lió, Carla Gomes, Max Welling, et al. 2022. Structure-based drug design with equivariant diffusion models. *arXiv preprint arXiv:2210.13695* (2022).

Vedant Singh, Surgan Jandial, Ayush Chopra, Siddharth Ramesh, Balaji Krishnamurthy, and Vineeth N Balasubramanian. 2022. On conditioning the input noise for controlled image generation with diffusion models. *arXiv preprint arXiv:2205.03859* (2022).

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*. PMLR, 2256–2265.

Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems* 32 (2019).

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

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

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. 2023. De novo design of protein structure and function with RFdiffusion. *Nature* (2023), 1–3.

Wikipedia, the free encyclopedia. 2012. Amino acid structure. (2012). https://commons.wikimedia.org/wiki/File:Amino_acid_generic_structure.png#/media/File:Amino_acid_zwitterions.svg [Online; accessed October 5, 2023].

Derek N Woolfson. 2021. A brief history of de novo protein design: minimal, rational, and computational. *Journal of Molecular Biology* 433, 20 (2021), 167160.

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

Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. 2023. Diffusion probabilistic modeling for video generation. *Entropy* 25, 10 (2023), 1469.

Lvmin Zhang and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543* (2023).

Mengchun Zhang, Maryam Qamar, Taegoo Kang, Yuna Jung, Chenshuang Zhang, Sung-Ho Bae, and Chaoning Zhang. 2023. A survey on graph diffusion models: Generative ai in science for molecule, protein and material. *arXiv preprint arXiv:2304.01565* (2023).

Qinsheng Zhang and Yongxin Chen. 2022. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902* (2022).