A Survey on Conditional Protein Generation Using Diffusion Models

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

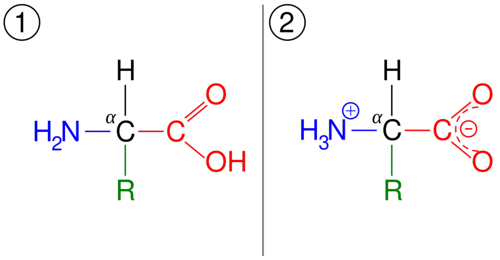

Figure 1: The molecular structure of an amino acid in its (1) un-ionized and (2) zwitterionic forms. Showing the central alpha carbon (black), the carboxyl group (red), the amino group (blue), and the variable side chain (green)[Wikipedia, the free encyclopedia 2012]

ABSTRACT

Abstract

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.

Given the pivotal role of proteins in living organisms, a crucial need for designing proteins arises, whether it be to enhance their activities or to create new ones. Unfortunately, designing proteins faces a major hurdle due to the enormity of the protein search space, which encompasses 20^{100} potential amino acid sequences for a 100-residue protein. Moreover, natural evolution has only explored a small fraction of this expansive space. 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, results in significant challenges in developing effective tools for designing *de novo* protein sequences with specific desired structures and features.

Historically, protein analysis relied on labour-intensive experimental techniques, which required significant expertise [Jaskolski et al. 2014]. The advent of computational methods and machine learning methods enabled a more efficient exploration of the protein search space and harnessed the capabilities of these new methods. 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. While VAEs can produce more 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 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. For the generation of novel proteins are these sets of features for diffusion models highly relevant, thus making them a pivotal tool,

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 designing entirely new proteins. The generation of proteins can involve generating desired sequences or structures either partial or complete. Sequence and structure co-design involves generating sequences and structures jointly.

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

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, in the context of this survey about protein design, refers to the deliberate introduction of additional information to guide the generative part of the diffusion process towards a specific, intended outcome. This desired outcome is thus influenced by the information provided. Instead of sampling a protein from the total learned distribution, conditioning ensures that a sample is sampled from a smaller intended subset of the total distribution. This conditioning information is a protein's properties such as the protein's own desired (partial) structures or sequences or other biophysical properties. This can also be characteristic information about a binding molecule's (partial) structure or sequences. Conditioning is thus a crucial element of protein design, as it allows designers to tailor protein properties to meet specific objectives.

Give an example

This survey begins by providing background on the foundational concepts of diffusion models in Section 2. Subsequently, we explore various conditional settings in Section **??**. Section 7 highlights recent advancements in related fields that could be applied to 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. By understanding the process of removing the noise, the model can generate novel outputs of the data distribution.

Three sub-types exist within this category: Denoising Diffusion Probabilistic Models (DDPMs), Score-based Generative Models (SGMs), and Stochastic Differential Equations (SDEs). 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.

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, utilising 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:

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

Lastly, the backward diffusion process uses a neural network θ that learns to predict the noise that was added in a forward step. This backward process is designed to reconstruct 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 conducted by [Song et al. 2020], the authors revealed the connection between Stochastic Differential Equations (SDEs) and Score-Based Generative Models (SGMs) along with the Generative Diffusion Models (DDPMs) 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 discretize 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 with respect 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 characterized as follows:

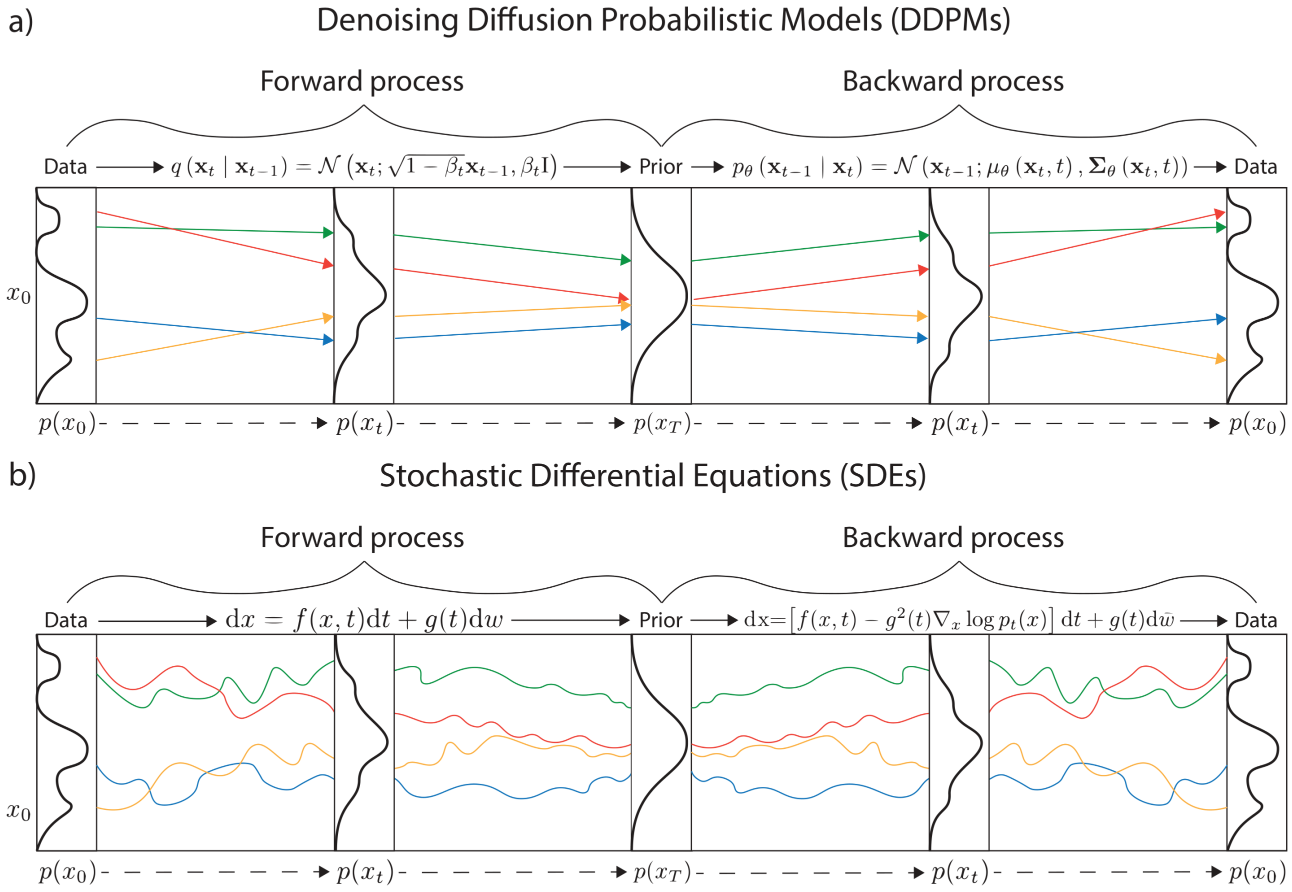
a)
Denoising Diffusion Probabilistic Models (DDPMs)
Forward process
Backward process
Data
$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathrm{I})$
Prior
$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \mathbf{\Sigma}_\theta(\mathbf{x}_t, t))$
Data
x_0
x_0
$p(x_0)$
$p(x_t)$
$p(x_T)$
$p(x_t)$
$p(x_0)$
b)
Stochastic Differential Equations (SDEs)
Forward process
Backward process
Data
$\mathrm{d}x = f(x,t)\mathrm{d}t + g(t)\mathrm{d}w$
Prior
$\mathrm{d}x = [f(x,t) - g^2(t)\nabla_x \log p_t(x)]\,\mathrm{d}t + g(t)\mathrm{d}\bar{w}$
Data
x_0
x_0
$p(x_0)$
$p(x_t)$
$p(x_T)$
$p(x_t)$
$p(x_0)$

Figure 2: Caption

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

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 further utilized 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 through parameterization in a score model, denoted as $s_\theta(xt,t)$. This process extends the score-matching objective to continuous time as specified by [Yang et al. 2023].

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

In summary, DDPM and SDE are distinct generative modelling techniques with different underlying principles and applications. DDPM focuses on autoregressive 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 you are working with.

2.2 Pre-Processing, Post-Processing and Validation

Pre-Processing, Post-Processing and Validation

3 DESIGNING PROTEINS BY CONDITIONED ON STRUCTURE

A protein's structure 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 intricate folded state of the entire amino acid chain; and, in cases where proteins consist of multiple amino acid sequences, the quaternary structure, determined by the arrangement of these chains.

Designing by conditioning on motifs or secondary structures can have several advantages. One motivation is that a motif plays a significant role in influencing the mechanical properties of protein materials. It is thus beneficial to condition on certain motifs to design proteins with these mechanical characteristics. Another reason is flexibility since this conditioning approach does not overly

Authors	Diffusion Type	Architecture	Generative type	Conditioning type	Generates
[Ni et al. 2023]	DDPM	UNET	De novo	CoStr	Sequence
[Trippe et al. 2022]	DDPM	EGNN	De novo and optimization	CoStr	Structure
[Anand and Achim 2022]	DDPM	ETNN	De novo and optimization	CoStr	**Structure**
[Jing et al. 2023]	HDM	EGNN	Structure prediction	CoSeq	Structure
[Qiao et al. 2022]	SDE	EGTNN	Structure prediction	CoSeq, CoBM	Structure
[Nakata et al. 2023]	DDPM	EGNN	Structure prediction	CoSeq, CoBM	Structure
[Martinkus et al. 2023]	DDPM	APMixer	De novo	CoBM	Co-design
[Luo et al. 2022]	DDPM	MLP	De novo and optimization	CoBM	Co-design
[Ketata et al. 2023]	SDE	ECNN	Structure prediction	CoBM	Structure
[Watson et al. 2023]	DDPM	RoseTTAFold	De novo and optimization	CoStr, CoSeq, CoBM	Structure

Table 1: Tabulated summary of reviewed papers. The initial column displays contributing authors. The next column highlights the Diffusion types (Harmonic Diffusion Model (HDM) is a new diffusion model developed by [Jing et al. 2023]). The third column specifies the architecture utilized in the backward diffusion process (Aligned Protein Mixer (APMixer) is a novel MLP architecture developed by [Martinkus et al. 2023]). The generative type column distinguishes between de novo protein generation, optimization of existing proteins, or structural prediction. Following that, the table indicates the varied conditioning types used, including Structure (CoStr), Sequence (CoSeq), or Binding Molecule (CoBM). The final column denotes whether the model primarily predicts sequences, predicts structures or applies a co-design approach.

constrain the model to produce only a single structure. It allows for significant variation in the generated protein structures while still adhering to the specified secondary structure constraints.

The works from [Anand and Achim 2022] proposed an equivariant DDPM for generating protein structure backbones conditioned to topological constraints. On top of these generated backbones, they can diffuse the protein's sequence and rotamers to obtain the full protein. Notable from their work is that their generated proteins do not need a post-processing step which makes their model end-to-end. Their obtained results show that their model can successfully do region recovery or optimization using inpainting. It also has comparable results when generating sequences and rotamers to 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. Co-designing the complete protein allows a model to have more qualitative results since it can cross-condition on structure, sequence, and rotamers.

The works from [Ni et al. 2023] suggested utilizing a DDPM to produce novel protein sequences while conditioning on secondary structure information. Their model differs from other approaches since they bypass the construction of atomic details of the backbone and only focus on the mapping between motifs and sequences. Their model generates sequences by conditioning on a fractional distribution over 8 different types of secondary structures. It can also take in more specific information in the form of per-residue secondary-structure information. After generating these novel sequences they use the folding prediction methods to predict the proteins' complete structures and classify their secondary structures. From their results where they compare these classifications with the input conditioning information, they can conclude that their novel proteins closely follow the conditioning information. Lastly, they validate the generated proteins' novelty by doing a BLAST analysis and find that giving the per-residue information is the most effective in generating de novo sequences (on the order of 50% – 60% similarity).

[Trippe et al. 2022]

4 PREDICTING STRUCTURE BY CONDITIONING ON SEQUENCE

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.

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. First [Qiao et al. 2022] proposed NeuralPLexer, a deep generative model that samples the dynamic protein-ligand structures conditioned on sequence and ligand molecular graph input. It utilizes a PLM to obtain protein features from the input sequence as well as AlphaFold2 to generate structural templates from the input sequence. Another diffusion-based model that generates dynamic protein-ligand structures was developed by [Nakata et al. 2023] that also utilized a PLM but removed the dependency for structural templates. While giving good results, this model has been limited by protein sizes and a small amount of training data. To address the shortcomings of both these models a novel model, Eigenfold [Jing et al. 2023] that generates dynamic protein structures was proposed. While it doesn't quite connect single-structure prediction and structural ensemble prediction, its results and method do set the groundwork for predicting dynamic structures.

[Trippe et al. 2022]

5 PROTEIN DESIGN CONDITIONED ON OTHER MOLECULES

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 are crucial for the development of novel therapeutics and biotechnological applications. This section will delve into the latest research focused on conditioning these molecular interactions.

5.1 Protein binding structure generation conditioned on ligand

In a diffusion setting, the works by [Qiao et al. 2022] were the first to cast the problem of finding protein structure given a ligand structure. A ligand is a molecule, often a small chemical compound, that binds to a protein's active site, modulating the protein's function or activity. Their novel model, NeuralPlexer, tested on ligand-binding proteins that exhibit large conformational variability exhibits the highest performance compared to the best-performing structure prediction methods. These other methods do not take into account ligand information, showing from their results that this information is necessary.

Research from [Watson et al. 2023] modelled protein-ligand interaction without structural information, unlike NeuralPlexer. During testing, they showcase that leaving out this structural information compared to including this information still results in properly generated structures. Next, they 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.

[Nakata et al. 2023] [Trippe et al. 2022]

Other notable research, such as DiffDock [Corso et al. 2022] and DiffBP [Lin et al. 2022], have 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 beyond the scope of this study.

5.2 Protein-protein interaction

Inspired by the DiffDock model [Ketata et al. 2023] 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 their 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 results.

5.3 Designing antibodies conditioned on antigens

Antibodies are specialized 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 specific 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.

The first to use a diffusion-based approach to design CDRs is from [Luo et al. 2022]. Their model, dubbed DiffAb, is explicitly conditional on the 3D structure of the antigen, allowing it to generalize to new antigens. Notably, DiffAb can co-design sequences and structures and optimize binding energy for existing antibodies, yielding competitive performance.

The work of [Martinkus et al. 2023] takes into account that large protein families typically have strong properties, such as being able to be mapped to a reliable sequence ordinate via sequence alignment. To this end, they developed AbDiffuser, which extends beyond CDR generation to encompass the full antibody structure. Generating complete antibodies broadens design possibilities such as optimizing stability or immunogenicity, and potentially impacting antigen interaction and CDR conformation.AbDiffuser demonstrates robust antibody generation with lower memory requirements compared to 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.

6 UNCONDITIONAL PROTEIN SAMPLING

unconditional Protein Sampling

7 FUTURE WORK

future work

8 CONCLUSION

conclusion

REFERENCES

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.

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).

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

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.

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).

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).

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).

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).

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.

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, 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).

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].

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.

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).