A Survey on Conditional Protein Generation Using Diffusion Models

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

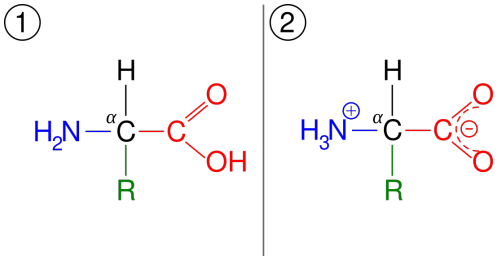

Figure 1: The molecular structure of an amino acid in its (1) un-ionized and (2) zwitterionic forms. Showing the central alpha carbon (black), the carboxyl group (red), the amino group (blue), and the variable side chain (green)[Wikipedia, the free encyclopedia 2012]

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 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 citation?, 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, demanding significant expertise [Jaskolski et al. 2014]. The advent of computational methods enabled a more efficient exploration of the protein search space. Furthermore, the rise of machine learning models across various fields has motivated scientists to harness their capabilities for navigating this complex space. 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 outputs. 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 outputs. While VAEs can produce more diverse outputs they often lack in 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. Moreover, diffusion models offer rotation-equivariant outputs. These sets of features for diffusion models are highly relevant, thus making them a pivotal tool, for the generation of novel proteins.

The generation of proteins involves creating new protein sequences or structures. There are two primary types of protein generation, either by optimizing existing proteins or creating novel proteins. Optimizing existing proteins involves refining or modifying existing protein structures to improve their properties, such as binding affinity. The creation of novel proteins requires designing entirely new proteins conditioned on specific functions or properties. The generation of proteins can involve generating desired

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

motifs, backbone structures, or complete structures, including backbone and sidechain placement. This can be done by targeting either the protein sequence, the protein structure or both.

Prior 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. Specifically, we delve into the generation conditioned on protein motifs, antibodies, structure, sequence, and protein-ligand interactions.

Give motivation on why we need conditioning

This survey begins by providing background on the foundational concepts of diffusion models in Section 2. We subsequently explore various conditional settings in Section 3. Section 4 highlights recent advancements in related fields that could be applied to protein design. Finally, we conclude this survey in Section 5.

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. Note: there is a fourth, harmonic diffusion, introduced in a recent protein structure generation paper by [Jing et al. 2023]

2.1.1 Denoising Diffusion Probabilistic Models(DDPM). A Denoising Diffusion Probabilistic Model (DDPM) is a type of generative model capable of creating new data samples from a specified data distribution, utilizing 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 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 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:

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

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 $p_\theta(x_{t-1}|x_t)$ the model's optimization 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 $\Sigma_\theta(x_t, t)$ (I know that for some CV applications, they do predict the variance but I am not sure if this is also done for Protein Design) predict the mean and the variance of the noise at time t respectively. 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 Score-based Generative Models (SGMs).

2.1.3 Stochastic Differential Equations (SDEs).

2.2 Protein Representation

There are two ways one can represent a protein. The conventional way is by using the 3D coordinates of all amino acid residues. The other technique, which is less frequently used, is describing the position of the next residue given the current residue.

The 3D Cartesian space representation technique Reference? represents a protein by 3D coordinates of all its amino acid residues. It considers the spatial arrangement of atoms in the protein. To handle rotational symmetries, equivariant deep learning models are used, which ensure that the model's output maintains rotational invariance. During the forward diffusion process, some random residues may be masked, and this masking is reversed during sampling.

The second representation technique uses a set of angles, called the dihedral angles, to describe the position of the next residue given the current residue. This allows for using bidirectional transformer architectures instead of complex equivariant models. However, converting the protein from an angle to a 3D coordinate representation can be challenging. Angle representations can lead to structural collisions and errors propagating through the protein structure, leading to a domino effect.

2.3 Pre- and Post-Processing

3 CONDITIONAL PROTEIN GENERATION

Instead of writing this section I have been mainly working on a mindmap with different categories on how to divide the papers that I have found. You can find it by following this link: https://www.figma.com/file/AMaKzminXcZ6fI4FhUQZpi/Mindmap-Diffusion-for-protein-design?type=whiteboard&node-id=0%3A1&t=1hupAOijlNdh1

3.1 Antibodies - type of protein

[Martinkus et al. 2023] [Luo et al. 2022]

3.2 Structure

[Anand and Achim 2022] [Yi et al. 2023]

3.2.1 Motifs or secondary structure. Use this paragraph to help explain structure Protein structures are classified into four levels: the primary structure, which denotes the amino acid sequence without 3D considerations; the secondary structure, defining localized folding patterns or motifs like 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.

3.3 Sequence

[Nakata et al. 2023] [Qiao et al. 2022] [Jing et al. 2023]

3.4 Protein-Ligand Interaction

[Corso et al. 2022] [Nakata et al. 2023] [Qiao et al. 2022] [Watson et al. 2023]

4 FUTURE WORK

ControlNet, a neural network architecture to add spatial conditioning controls to large, pretrained text-to-image diffusion models. This allows the model to use for example edge maps to control the final output of the model [Zhang and Agrawala 2023].

5 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).

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.

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).

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).

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.

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.

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].

Kai Yi, Bingxin Zhou, Yiqing Shen, Pietro Liò, and Yu Guang Wang. 2023. Graph denoising diffusion for inverse protein folding. *arXiv preprint arXiv:2306.16819* (2023).

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).