Preprint: Under Review 1–13

Controllable Flow-based Protein Generation

Firstname Lastname NAME@EMAIL.EDU
Department
University
City, State, Country

Abstract

Generative models are increasingly vital in protein design, yet conditional generation for specific targets remains challenging. Current methods often require extensive screening and lack fine-grained control over functional traits. We propose a flow-matching-based multimodal model that jointly generates protein sequences and structures conditioned on a target. Our approach integrates sequence and structure modalities using a locality-aware transformer with self-attention and cross-attention mechanisms. Experimental results on fine-grained protein generation, shows significant improvements over baseline methods.

1. Introduction

Generative models are playing increasingly important in biological sciences (Guo et al., 2024), particularly in protein design (Abramson et al., 2024). While typical protein design models often focuses on creating novel proteins, therapeutic applications frequently require designing proteins for specific targets (Nelson et al., 2010). This necessitates conditional protein generation: that is models that can predict or generate proteins conditional on satisfying target characteristics. Despite the success of models like EvoDiff (Alamdari et al., 2023), ProteinGAN (Repecka et al., 2021), MSAVAE (Hawkins-Hooker et al., 2021), these protein design models have limited practical benefit as researchers need to screen the generated proteins to fulfill the desired criteria (Mardikoraem et al., 2023).

Recent innovations in generative modeling have addressed some of these limitations. TaxDiff (Zongying et al., 2024) guides generation via species taxonomy, though this fails to specify functional traits. CMAdiff (Zhou et al., 2025)addresses this by integrating CVAE with diffusion models for function-driven design. However these can only leverage very high level information, and do not work well for tasks involving more fine-grained control. Vázquez Torres et al. (2024) demonstrated a deep learning-based method, to design high-affinity binders. Similarly, Wu et al. (2024)developed approaches to design binders for intrinsically disordered regions (IDRs) of proteins.

Despite these advancements, challenges remain. Many of these methods are focused on specific targets. Furthermore designed proteins may adopt unintended conformations, affecting their functional utility. Addressing these issues requires models that consider both the amino acid sequence and the three-dimensional structure of proteins.

Contributions We present a flow-matching (Lipman et al., 2022) based generative model that simultaneously generates protein sequences and their corresponding structures, conditioned on a target protein. Following Yim et al. (2023b) we propose a multimodal

generative model that learns factorized flows for each data modality. We achieve this by using a locality aware multi-modal transformer model that integrates information from both modalities (sequence and structure) to learn the flow. Our method combines full self-attention blocks (Vaswani et al., 2017) with bottlenecked cross attention blocks (Nagrani et al., 2021) and distance sensitive attention layers to that can capture both global dependencies as well as local interactions.

2. Preliminaries

2.1. Flow Matching

Flow Matching (FM) Lipman et al. (2022) is a simulation-free framework for learning an ordinary differential equation (ODE) of the form:

$$\frac{dx_t}{dt} = v(t, x_t),$$

where the velocity field v is parameterized by a neural network v^θ. The goal is to learn a transformation from a source distribution p_0 to a target distribution p_1. The flow of functions ψ_t induced by this velocity field generates a time-indexed density path p_t via the pushforward $p_t = [\psi_t]_\# p_0$. This evolving distribution satisfies the continuity equation: $\frac{\partial p_t(x_t)}{\partial t} = -\nabla_{x_t} \cdot (p_t(x_t) u_t(x_t))$, where $u_t(x_t)$ is the (generally intractable) true velocity field associated with the flow.

Conditional Flow Matching (Lipman et al., 2022), define a conditional probability path $p_t(x_t \mid z)$ conditioned on a latent variable z, such that the conditional vector field $u_t(x_t \mid z)$ becomes tractable. The latent variable z is chosen to simplify the construction of valid sample paths.

A practical instantiation (Lipman et al., 2022) is to let z depend on $x_1 \sim p_1$, and define a simple interpolation (e.g., linear) that ends at x_1. This gives rise to known intermediate distributions $p_t(x_t \mid z)$ and an explicit formula for the conditional velocity field $u_t(x_t \mid z)$.

The learning objective becomes:

$$\mathbb{E}_{z,t,x_t \sim p_t(x_t|z)} \left[\left\| v^\theta(t, x_t) - u_t(x_t \mid z) \right\|^2 \right].$$

Lipman et al. Lipman et al. (2022) show that this conditional objective correctly recovers the marginal velocity field $u_t(x_t)$ when integrated over the conditioning variable z, ensuring the learned dynamics transport p_0 to p_1 as desired. Since then, more generalized variants of this problem have been proposed (Tong et al., 2023). Any suitable conditioning variable z can be chosen if the objective remains tractable (Pooladian et al., 2023; Tong et al., 2023).

2.2. Transformers

Transformers Vaswani et al. (2017) are neural architectures that rely on **self-attention** to process sequential data without recurrent or convolutional operations, and have become the de-facto models in deep-learning applications. A single transformer block consists of two core components: a self-attention layer SA and a feedforward networks FFN stacked with residual connections and norm operations. Given an input sequence of embeddings

$\mathbf{X} \in \mathbb{R}^{n \times d}$ (for n tokens of dimension d), SA first computes Query, Key, Value Projections (denoted as Q, V, K). Q of a token i and K of all tokens are used to weigh the different input tokens, and then these weights are used to aggregate the V projections.

$$\mathbf{Q} = \mathbf{X}W_Q, \quad \mathbf{K} = \mathbf{X}W_K, \quad \mathbf{V} = \mathbf{X}W_V, \qquad \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$$

One can further define cross-attention between two inputs $\mathbf{X}, \mathbf{Y}$ denoted as $(MCA(\mathbf{X}; \mathbf{Y}))$ by forming queries from $\mathbf{X}$ and keys and values from $\mathbf{Y}$.

Multiple such attention heads (either self or cross) are concatenated and then passed to a position wise MLP and then combined as follows:

$$\mathbf{X}_{l+1} = Trans(\mathbf{X}_l) = \begin{cases} \mathbf{X}_{l-0.5} = \mathbf{X}_{l-1} + \text{SA}_l(LayerNorm(\mathbf{X}_{l-1}))) \\ \mathbf{X}_l = \text{FF}_l(LayerNorm(\mathbf{X}_{l-0.5})) + \mathbf{X}_{l-0.5} \end{cases}$$

2.3. Other Related Work

Models like AlphaFold2 (Jumper et al., 2021; Evans et al., 2021) and RoseTTAFold (Board, 2021) have significantly improved the accuracy of protein structure prediction (Bennett et al., 2023). RFdiffusion, developed by Watson et al. (2023), fine-tunes the RoseTTAFold (Board, 2021) network to achieve high-performance protein backbone generation, facilitating the design of protein monomers, binders, and symmetric oligomers. These models employ denoising diffusion (Ho et al., 2020) , initially developed for image generation, to iteratively refine protein structures from random noise. Recently, we have also seen the emergence of models like AlphaProteo (Zambaldi et al., 2024) and Chroma Ingraham et al. (2023) which are generative model capable of producing novel protein structures and sequences, conditioned on desired properties and functions. Vázquez Torres et al. (2024) and Wu et al. (2024) have developed models specifically focused on producing better binding proteins. Yet, despite these advancements, challenges remain in the conditioned design of protein sequences and structures (Zhou et al., 2025).

FrameDiff (Yim et al., 2023a), utilizes diffusion on the SE(3) group to model protein structures, enabling designable without relying on pretrained structure prediction networks. Building upon Yim et al. (2023a), FoldFlow (Bose et al., 2024), employs a flow-matching paradigm over 3D rigid motions, enhancing the modeling power for protein backbones. FrameFlow (Yim et al., 2023b), an adaptation of FrameDiff, leverages SE(3) flow matching to achieve faster and more efficient protein backbone generation. More recently Geffner et al. (2025) introduced a large scale flow matching model for structure generation.

MultiFlow (Campbell et al., 2024), introduced a multimodal flow-based modeling framework that simultaneously generates protein sequences and structures, achieving state-of-the-art co-design performance. Genie(Lin and AlQuraishi, 2023), presents a diffusion model that generates protein backbones as sequences of C_α atomic coordinates, performing diffusion directly in Cartesian space and utilizing SE(3)-equivariant denoising.

3. Method

Following the design of Campbell et al. (2024), we represent a protein as a sequence and structure of a chain of amino acids. Let $\mathcal{A}$ be the set of 20 common amino acids; then a

protein of length L is considered as consisting of two modalities: a) a sequence of L amino acids $\mathbf{x}^A = a_1, a_2, ...a_L$ and b) the 3D coordinates of the amino acids. Following prior works Lin et al. (2023b); Lin and AlQuraishi (2024); Geffner et al. (2025), we model protein residue locations using their 3D α-carbon (C_α) coordinates. Specifically, the backbone coordinates are given by a vector $\mathbf{x}^R \in \mathbb{R}^{3L}$. Given a target function, the goal of the model is to predict a corresponding protein (for example a protein to bind to an antigen), in the form of its amino acid sequence and their alpha carbon coordinates.

3.1. Training

We propose a multimodal factored flow matching approach to generate the binding protein. The flow model used in our approach follows the basic design in Campbell et al. (2024) but takes a conditional embedding c as additional input. Specifically, we have two velocity models v^R and v^A, both of which are conditioned on the inputs x_t^R, x_t^A but produce independently the velocities in each modality i.e.

$$v^\theta(x_t, t; c) = [v^{\theta,A}(x_t, t; c), v^{\theta,R}(x_t, t; c)]$$

where x_t represents the joint modalities x_t^R, x_t^A. During sampling the model computes the velocity field, and updates the two modalities separately. The details on how the two modalities x_t^R, x_t^A are fused into a single representation x_t and the architecture of model for v^θ are discussed in the next section.

Other than the flow matching objective, we also provide additional structural supervision using the distogram loss(Abramson et al., 2024; Qu et al., 2024). This is computed as the binned pairwise distances $D_{ij}^{(b)}(x)$ between residues i and j, where $b \in \{1, \ldots, 64\}$ indexes the discrete bins (Abramson et al., 2024; Qu et al., 2024; Geffner et al., 2025). The distogram loss is computed via a prediction head attached to the architecture's pairwise representation.

During training, we adopt the standard noisy linear interpolant scheme $x_t = tx + (1-t)x_0 + \epsilon$ used in flow-matching (Lipman et al., 2022), and define the model's objective as a combination of a flow-matching loss and the distogram prediction loss:

$$\min_\theta \; \mathbb{E}_{t,c,x_0,x_1,\epsilon}\left[\underbrace{\left\|v_t^\theta\left(x_t, t; c, \hat{x}(x_t)\right) - (x_1 - x_0)\right\|_2^2}_{\text{FM loss}} - \sum_{i,j}\sum_{b=1}^{64} D_{ij}^{(b)}(x) \log p_{b,ij}^\theta\left(x_t, t, \hat{x}(x_t)\right)\right] \tag{1}$$

Additionally, following the self-conditioning protocol of Geffner et al. (2025) we introduce $\hat{x}(x_t)$ as an additional input to the velocity model; and is an estimate of the clean data:

$$\hat{x}(x_t) = x_t + (1-t) \cdot v_t^\theta\left(x_t, t; c\right)$$

During training the self-conditional $\hat{x}$ is dropped or kept with probability 0.5 for each sample.

3.2. Representation

Residue Representation While discrete token-level models have been explored for flow-matching (FM), continuous representations are generally more compatible with FM-based generative methods. However, to enable effective cross-modal modeling, these continuous representations must also retain semantic meaning relevant to the underlying protein sequence. To this end, we leverage the latent space of a pretrained protein language model (pLM), following the approach of Hu et al. (2024); Kong et al. (2025).

High-dimensional embeddings from pLMs can present challenges, including poor scalability and degraded generative performance, particularly in low-data regimes. To address this, we project the high-dimensional latent representations into a lower-dimensional, semantically compressed space. Specifically, we use the pretrained ESM-2 model Lin et al. (2022) as the base encoder and insert a single transformer layer to project the token embeddings into a reduced latent space. A corresponding output layer is added as the first layer in the ESM-2 decoder for sequence reconstruction.

We employ a two-stage training strategy: first, we perform LoRA fine-tuning of the ESM-2 encoder and decoder on the training data. Once adapted, these components are frozen, and only the newly added projection and reconstruction layers are fine-tuned. The encoder maps the protein sequence $\mathbf{x}^A = [a_1, a_2, \ldots, a_L] \in \mathcal{A}^L$ to a continuous representation $X = [x_1, x_2, \ldots, x_L] \in \mathbb{R}^{L\times D}$. This is then further downsampled into a lower dimensional vector $\in \mathbb{R}^{L\times d}$ by a transformer block. This is the latent representation that is used as training data for the flow matching model. The explicit sequence is reconstructed by the decoder during sampling and for distogram computation.

Coordinate Representation Since the coordinates are real numbers, flow matching can usually deal with it effectively. Recent works (Bose et al., 2024) have proposed using SE(3)-group based representations to capture the rigid "frames" (Jumper et al., 2021). These then use non-Euclidean manifolds (and non-linear interpolants) to model the flow. Combining these with the euclidean latent space of sequence can be challenging (Huguet et al., 2024; Geffner et al., 2025) and hence we simply used linear coordinates.

3.3. Network Architecture

Let $X \in \mathbb{R}^{L\times d}$ denote an input sequence of L tokens with embedding dimension d. For simplicity we will only refer to the sequence modality as X, but the architecture for the structure/coordinate modality is identical. We use transformer based neural network to learn the velocity field. However along with standard vanilla transformer based encoder, we also include a locality-aware layer to specifically model local interactions and a bottleneck fusion layer to incorporate cross-modal attention. Our overall network architecture comprises three main components:

Multi-layer Self-Attention We apply N layers of standard multi-head self-attention and feed-forward sublayers:

$$X^{(0)} = X, \quad H^{(\ell)} = Trans_\ell(X^{(\ell-1)})$$

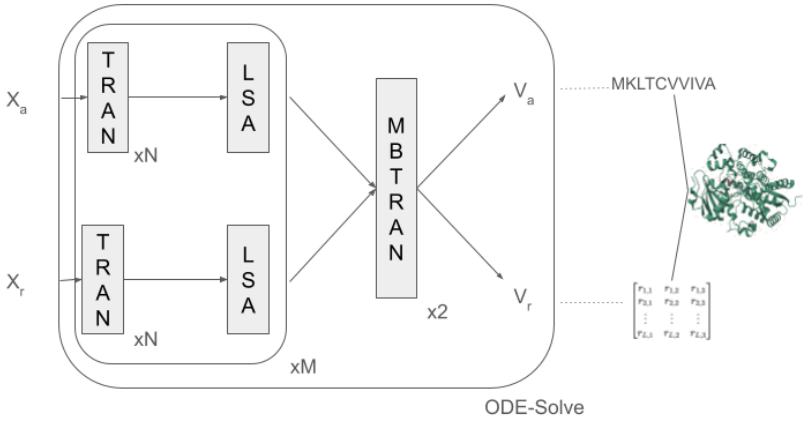
Locality-Sensitive Attention (LSA) (Zongying et al., 2024) found that patching attention to adhere to spatial locality improves generation. We also incorporate an additional

locality sensitive transformer block which is just an ordinary transformer block bu for which the attention mechanism has been modified to pay greater emphasis on spatially or structurally nearby tokens. Let $D \in \mathbb{R}^{L \times L}$ be a matrix of pairwise distances, then the attention weights are computed as:

$$\alpha_{ij} = \frac{\exp\left(\frac{q_i^\top k_j}{\sqrt{d_k}} \cdot \phi(D_{ij})\right)}{\sum_j' \exp\left(\frac{q_i^\top k_{j'}}{\sqrt{d_k}} \cdot \phi(D_{ij'})\right)} \qquad \text{LA}(H)_i = \sum_{j=1}^{L} \alpha_{ij} v_j$$

where $\phi(D_{ij})$ is a gaussian weighting kernel.

Figure 1: Model architecture for the Flow Model.



Multi-Modal Bottleneck Fusion Full cross-attention between modalities is both slow to compute and can be less performant, especially when training data is small (Nagrani et al., 2021). To alleviate this we introduces multimodal fusion bottleneck tokens (Nagrani et al., 2021) in each modality, and all cross-modal attention flow is restricted to be via these fusion bottleneck tokens. Let $Z \in \mathbb{R}^{B \times d}$ denote B bottleneck tokens, and $Y \in \mathbb{R}^{L' \times d}$ represent another modality (e.g., structure, image). The overall fused representation is then given by $Z = Z_X + Z_Y$. The information from this fused representation is obtained by re-integrating Z into the token sequence using standard self-attention. More precisely cross-modal bottleneck attention between X and Y is given by:

$$X_{l+1}, Z_{l+1}^X = Trans(X_l || Z_l) \quad Y_{l+1}, Z_{l+1}^Y = Trans(Y_l || Z_l) \quad Z_{l+1} = Z_{l+1}^X + Z_{l+1}^Y$$

We take $X = X^A, Y = X^R$ as the sequence and structure modalities.

Full Network The overall network uses N layers for self-attention, followed by one block of transformer with locality-aware attention. This is followed by a block of bottleneck cross-modal attention. This constitutes an overall unit, and the overall network comprises of stacked layers of the described units as presented in Figure 1.

4. Experiments

We evaluated our proposed method on two representative tasks: a) to create proteins based on text-based annotations of the protein including information like organism taxonomy and high-level terms related to protein properties and b) sampling binder proteins for an antigen-antibody complex design.

4.1. Text Based Generation

We first focus on the task of text-conditioned generation where the problem is to sample proteins based on a certain desired character given as a text.

Training The data is curated from the Swiss-Prot dataset which, along with proteins, has additional expert annotations including function, domain structure, etc. We follow the protocol of Zhou et al. (2025) which paired protein sequences i) taxonomic information about the organism, ii) comments about function and domain structure, iii) high-level keywords about properties such as ribosomal protein, membrane protein etc. The conditioning variable is obtained as the embedding of the prompt text. For comparability with earlier works we use embeddings obtained from MiniLM-L6-v2.

Baselines Following earlier literature Zhou et al. (2025); Zongying et al. (2024), we compare against the following conditional generation approaches as baselines CARP (Yang et al., 2024), ProtGPT2 (Ferruz et al., 2022), EvoDiff (Alamdari et al., 2023), TaxDiff (Zongying et al., 2024), and CMADiff (Zhou et al., 2025). Among these CARP and ProtGPT2 are autoregressive models, while the rest are diffusion based.

Evaluation We compare the models on the following metrics: a) pLDDT (predicted local distance difference) derived from AlphaFold-3 (Jumper et al., 2021; Abramson et al., 2024), measures the confidence of predicted structures ; b) pTM (template modeling score) which measures topological similarity between the generated protein and the closest match in known databases (Zhang and Skolnick, 2005); and c) Fident (fold identity score) (van Kempen et al., 2022) that measures the global structural identity between the generated sequence and the closest natural homologous protein (homolog).

Table 1: Model performance on text condition protein design task. (↑): higher is better, (↓): lower is better. Metrics are calculated with 1000 samples generated from each model.

Model	pLDDT ↑	pTM (%) ↑	Fident (%) ↑
CARP	44.5	39.8	11.9
LRAR	47.2	38.2	15.8
ProtGPT2	53.3	40.4	11.5
Evodiff	50.7	48.3	14.6
Taxdiff	64.3	49.1	17.8
CMADiff	68.7	52.0	17.4
Ours	70.1	52.4	18.1

4.2. Antibody Complex Design

Next we evaluate our model on the significantly more challenging problem of protein binder design. Specifically we focus on generating antibodies that bind to a given antigen. Unlike the previous section where the conditioning variable is an embedding of the input text, the conditioning variable in this task is the antigen protein. The conditioning variable C for the input protein is obtained from ESM2.

Baselines To the best of our knowledge there are no explicit models which focus on the antibody binding task. However ProteinMPNN (Dauparas et al., 2022) which learns to predict protein sequences based on a given structure can be modified to predict binder based on a protein sequence. Similarly, the more recent CMADiff (Zhou et al., 2025)model which is originally trained to condition only on text embedding, can be modified to take ESM-2 (Lin et al., 2023a) based embedding of a protein to predict the binder. We consider these as baseline models and compare them against our proposed method.

Training For this task we follow the methodology of Jin et al. (2021). We extract 4000 protein complexes from the Structural Antibody Database (Dunbar et al., 2014). Since this dataset is quite small, training a complex generative model from scratch was not fruitful. Hence we first obtained a 300k chain pair interfaces from the training of AlphaFold3 (Abramson et al., 2024). We use each chain in the pair as the conditioning variable to generate the other pair. Once pre-trained on this set, we then finetune on the antibody task.

Evaluation We assess designed protein-protein complexes using the following metrics a) pLDDT (predicted local distance difference) derived from AlphaFold (Abramson et al., 2024) b) ipTM (interface predicted TM-score) which evaluates inter-chain interactions (Abramson et al., 2024); and c) pTM estimates global structure accuracy (Zhang and Skolnick, 2004).

Results are presented in Table 2, from which we can see that our approach outperforms baseline models.

Table 2: Model performance on text condition protein design task. All metrics are positive (higher is better) metrics. Our method demonstrates superior performance, achieving the highest performance. Metrics are calculated with 1000 samples generated from each model.

Model	pLDDT ↑	pTM (%) ↑	ipTM ↑
ProteinMPNN	80.5	64.7	0.496
CMADiff	84.7	67.3	0.541
Ours	87.1	71.0	0.579

5. Conclusion

In this paper, we present a flow-matching based generative model for controllable protein generations. Our approach leverages ideas from Yim et al. (2023b), Geffner et al. (2025), and Zongying et al. (2024) to create a controllable flow matching based model. The model architecture is based on combining multi-modal bottleneck transformer and locality-sensitive attention with factorized flow models to simultaneously generate both sequence and structure. On both text-based generation and conditioned antibody design our model outperforms baselines.

References

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016): 493–500, 2024.

Sarah Alamdari, Nitya Thakkar, Rianne van den Berg, Neil Tenenholtz, Bob Strome, Alan Moses, Alex Xijie Lu, Nicolo Fusi, Ava Pardis Amini, and Kevin K Yang. Protein generation with evolutionary diffusion: sequence is all you need. *BioRxiv*, pages 2023–09, 2023.

Nathaniel R Bennett, Brian Coventry, Inna Goreshnik, Buwei Huang, Aza Allen, Dionne Vafeados, Ying Po Peng, Justas Dauparas, Minkyung Baek, Lance Stewart, et al. Improving de novo protein binder design with deep learning. *Nature Communications*, 14(1):2625, 2023.

Advisory Board. Rosettafold: Accurate protein structure prediction accessible to all. *Institute for Protein Design*, 2021.

Joey Bose et al. Se(3)-stochastic flow matching for protein backbone generation. *arXiv preprint arXiv:2310.02391*, 2024.

Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. *arXiv preprint arXiv:2402.04997*, 2024.

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

James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M Deane. Sabdab: the structural antibody database. *Nucleic acids research*, 42(D1):D1140–D1146, 2014.

R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper, and D. Hassabis. Protein complex prediction with alphafold-multimer. *bioRxiv*, 2021. doi: 10.1101/2021.10.04.463034.

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

Tomas Geffner, Kieran Didi, Zuobai Zhang, Danny Reidenbach, Zhonglin Cao, Jason Yim, Mario Geiger, Christian Dallago, Emine Kucukbenli, Arash Vahdat, et al. Proteina: Scaling flow-based protein structure generative models. *arXiv preprint arXiv:2503.00710*, 2025.

Zhiye Guo, Jian Liu, Yanli Wang, Mengrui Chen, Duolin Wang, Dong Xu, and Jianlin Cheng. Diffusion models in bioinformatics and computational biology. *Nature reviews bioengineering*, 2(2):136–154, 2024.

Alex Hawkins-Hooker, Florence Depardieu, Sebastien Baur, Guillaume Couairon, Arthur Chen, and David Bikard. Generating functional protein variants with variational autoencoders. *PLoS computational biology*, 17(2):e1008736, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, volume 33, pages 6840–6851, 2020.

Vincent Hu, Di Wu, Yuki Asano, Pascal Mettes, Basura Fernando, Björn Ommer, and Cees Snoek. Flow matching for conditional text generation in a few sampling steps. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 380–392, 2024.

Guillaume Huguet, James Vuckovic, Kilian Fatras, Eric Thibodeau-Laufer, Pablo Lemos, Riashat Islam, Cheng-Hao Liu, Jarrid Rector-Brooks, Tara Akhound-Sadegh, Michael Bronstein, et al. Sequence-augmented se (3)-flow matching for conditional protein backbone generation. *arXiv preprint arXiv:2405.20313*, 2024.

John Ingraham et al. Illuminating protein space with a programmable generative model. *Nature*, 620(7973):687–695, 2023. doi: 10.1038/s41586-023-06728-8.

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

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

Zitai Kong, Yiheng Zhu, Yinlong Xu, Hanjing Zhou, Mingzhe Yin, Jialu Wu, Hongxia Xu, Chang-Yu Hsieh, Tingjun Hou, and Jian Wu. Protflow: Fast protein sequence design via flow matching on compressed protein language model embeddings. *arXiv preprint arXiv:2504.10983*, 2025.

Yeqing Lin and Mohammed AlQuraishi. Generating novel, designable, and diverse protein structures by equivariantly diffusing oriented residue clouds. *Proceedings of the 40th International Conference on Machine Learning*, 202:1–10, 2023.

Zeming Lin and Mohammed AlQuraishi. Generative protein design with language models and diffusion. *bioRxiv*, 2024.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022:500902, 2022.

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

Zeming Lin, Hattie Akin, Roshan Rao, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. In *ICML*, 2023b.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2022.

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

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems*, 34:14200–14213, 2021.

Aaron L Nelson, Eugen Dhimolea, and Janice M Reichert. Development trends for human monoclonal antibody therapeutics. *Nature reviews drug discovery*, 9(10):767–774, 2010.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky TQ Chen. Multisample flow matching: Straightening flows with minibatch couplings. 2023.

Wei Qu, Jiawei Guan, Rui Ma, Ke Zhai, Weikun Wu, and Haobo Wang. P (all-atom) is unlocking new path for protein design. *bioRxiv*, pages 2024–08, 2024.

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

Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian FATRAS, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023.

Michel van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Foldseek: fast and accurate protein structure search. *Biorxiv*, pages 2022–02, 2022.

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

Susana Vázquez Torres, Philip JY Leung, Preetham Venkatesh, Isaac D Lutz, Fabian Hink, Huu-Hien Huynh, Jessica Becker, Andy Hsien-Wei Yeh, David Juergens, Nathaniel R

Bennett, et al. De novo design of high-affinity binders of bioactive helical peptides. *Nature*, 626(7998):435–442, 2024.

Joseph L Watson et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7973):687–695, 2023. doi: 10.1038/s41586-023-06415-8.

Kejia Wu, Hanlun Jiang, Derrick R Hicks, Caixuan Liu, Edin Muratspahić, Theresa A Ramelot, Yuexuan Liu, Kerrie McNally, Amit Gaur, Brian Coventry, et al. Sequence-specific targeting of intrinsically disordered protein regions. *bioRxiv*, pages 2024–07, 2024.

Kevin K Yang, Nicolo Fusi, and Alex X Lu. Convolutions are competitive with transformers for protein sequence pretraining. *Cell Systems*, 15(3):286–294, 2024.

Jason Yim et al. Se(3) diffusion model with application to protein backbone generation. *arXiv preprint arXiv:2302.02277*, 2023a.

Jason Yim et al. Fast protein backbone generation with se(3) flow matching. *arXiv preprint arXiv:2310.05297*, 2023b.

V. Zambaldi, D. La, A. E. Chu, H. Patani, A. E. Danson, J. Wang, and D. Baker. De novo design of high-affinity protein binders with alphaproteo. *arXiv preprint arXiv:2409.08022*, 2024.

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4): 702–710, 2004.

Yang Zhang and Jeffrey Skolnick. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic acids research*, 33(7):2302–2309, 2005.

Changjian Zhou, Yuexi Qiu, Tongtong Ling, Jiafeng Li, Shuanghe Liu, Xiangjing Wang, Jia Song, and Wensheng Xiang. Cmadiff: Cross-modal aligned diffusion for controllable protein generation, 2025.

Lin Zongying, Li Hao, Lv Liuzhenghao, Lin Bin, Zhang Junwu, Chen Calvin Yu-Chian, Yuan Li, and Tian Yonghong. Taxdiff: Taxonomic-guided diffusion model for protein sequence generation, 2024.

Appendix A. Additional Details

The ProtSemantic/Swiss-Prot dataset is directly availble at https://huggingface.co/sanyier312/PhysChemDiff/tree/main