Function-Controllable Protein Generation: Methods and Evaluation

Junyuan Zhou *

Department of Faculty of Science and Engineering, University of Nottingham Ningbo China, Ningbo, 315100, China

* Corresponding Author Email: scyjz31@nottingham.edu.cn

Abstract. Deep learning promotes design of proteins from simply foldable to operationally functional. It propels the generative design with programmable functionality in synthetic biology and drug discovery. This article presents three research methods. FoldingDiff performs the diffusion modelling in folding angle space, where complex topologies can be generated and the burden of equivariance alleviated. LigandMPNN method can accurately model 3D ligand surroundings. Therefore, it can obtain high sequence recovery rate and interface metrics on the specifically selected test sets. ProDomino method predicts insertion sites and can make verified uses light or small molecule-controlling switches in a lot of systems. All methods have restrictions. The essay has explored solving methods such as use the ProteinMPNN method to advance second stage screening. By furthering the physical validation and the expansion of complexed ligands, hereby indicates, with the following additional methods the physical validation and ligand complexes expansion. PinMyMetal method enables possible high-precision of transition metal pockets with the high-precision. The DIP-seq method can locate insertion hotspots with high tolerance. The boundaries between foldable proteins and the programmable functional proteins are closing with all improvements. As shown, this means that it can provide reproducible and comparable engineering strategies for drug development, synthetic biology and materials design.

Keywords: Protein design, Diffusion models, Ligand conditioning, Domain insertion, Functional controllability.

1. Introduction

Protein design isn't just about getting the right folding condition. It also has the goal of adding function to proteins. This can be credited to deep learning technology. After the breakthroughs in the structure of AlphaFold and the RoseTTAFold method, research attention has spread to more complexes, switchable states, and programmable functions. There is a great deal of enthusiasm in drug discovery and synthetic biology, which has increased the need for generative design methods that can control functionality. It is very important to consider the active site geometry, electrochemical features of ligand pockets, conformational dynamics, and regulatory elements at the same time when encoding function into proteins. However, purely geometric hard constraints rarely produce ideal results. Sequence-centric or hybrid paradigms have been shown to be more reliable when there is a shortage of structural data. Currently, in the research field of protein design, there is still a lack of a unified evaluation system and a reproducible experimental feedback loop.

There are still difficulties in horizontal comparisons, and the directions for further improvement are often unclear. These are the main problems that stop practical application. Representative studies show that the way to encode function is starting to appear, but it's far from perfect to make the functions work. The FoldingDiff method not only reduces the strict requirements of equivariance but also avoids chirality errors. It can handle complex topologies successfully. Instead of using Cartesian point clouds, it achieves this effect by doing diffusion in the space of folding angles. However, after generation, it requires subsequent relaxation to solve local atomic clashes. This issue shows that physical consistency constraints are still necessary [1]. The LigandMPNN method adds non-protein atomic contexts like small molecules, nucleic acids, and metals directly into the design process. It also gets high sequence recovery and ligand-reconstruction metrics on carefully selected test sets [2]. The ProDomino method focuses on predicting domain insertion sites. It can also quickly build single-

component protein switches that can be controlled by light or small molecules. These switches include members of the CRISPR-Cas family [3]. They are tested in both human cells and bacteria. The PoseBusters method includes geometric and chemical consistency checks. It also has docking feasibility assessments for evaluation and physical authenticity. By doing this, it shows that many AI-generated poses seem like real poses but aren't actually physically true. This shows that we need standardized benchmarks [4]. The uMol method can predict protein-small molecule complexes right from the sequence. Also, when we use it with PoseBusters-style screening at the right pLDDT thresholds, uMol can greatly increase the number of credible poses. This gives us a way to do functional screening and expand data even without crystallographic structures [5].

This essay focuses on the function-controllable generation of protein. In Section 2, it describes the mechanisms and application scenarios in structure-diffusion generation, ligand-conditioned sequence design and domain-insertion switches. In Section 3, it lists 4 main problems. They are designability shrinkage, topological bias, limited test-set coverage and high failure rates of insertion sites. Then it suggests some improvements. For example, it suggests a ProteinMPNN-based secondary screening [2], the uMol method combined with strict physical checks, and the DIP-seq method to find high-tolerance insertion hot zones [6]. The article also gives recommendations on evaluation metrics and experimental feedback. At the same time, it discusses the prospects and risk boundaries.

2. Overview of Major Approaches

2.1. FoldingDiff

The aim of the structure-diffusion generation methods is to encode protein functionality directly into the generative process. FoldingDiff method can explore new folds and conformations and also significantly improve the success rate of generating complex topologies by performing diffusion in folding-angle space. This diffusion model operates in the space of torsion angles instead of constructing protein structures in Cartesian coordinate point clouds. According to the original report article, under the scTM ≥ 0.5 threshold, 177 of 780 design targets have achieved structures which are close to native or designed models [7].

2.2. LigandMPNN

On a given scaffold or ligand pose, the method of LigandMPNN designs functionally relevant sequences conditioned. It can produce sequences which not only fold correctly but also exhibit targeted binding affinity and specificity by explicitly incorporating the three-dimensional structure and spatial orientation of the ligand. For instance, among its curated test sets, authors report a 77.5% sequence-recovery rate for metal-binding sites and 63.3% for receptor–ligand interfaces. In addition, it enables atomic-context conditioning yielding superior sequence and interface metrics and multiple experimental validations of both structure and function [8].

2.3. ProDomino

ProDomino method can also encode function into proteins. It is a machine-learning method for predicting permissible domain-insertion sites. The algorithm of this method first identifies positions in a target protein's sequence and then structure that tolerate insertion of an exogenous regulatory domain. Thus, it enables controllable protein switches. In both E. coli and human cells, the authors rapidly engineered CRISPR-Cas switches responsive to either light or small molecules and validated them by using ProDomino method [5]. For example, the Cas12a-LOV2 light-switch exhibited activity reduced to roughly one-third of that in darkness when under illumination, which demonstrates light-controlled switching. What's more, the editing efficiency recovered to about 70% of wild-type levels in the presence of cortisol when the glucocorticoid receptor domain GR2 was inserted into Cas12a-LOV2. Nevertheless, the efficiency dropped near the detection limit in its absence. Such obvious differences in editing activity have clearly illustrated the controllability of protein function [3].

3. Discussion

3.1. Challenges

Every method faces significant challenges. For the FoldingDiff method, the proteins which are generated by it exhibit further shrinkage in designability, topological diversity bias and atomic clashes. Only 20.9% of structures were judged designable after 177 qualified structures were assessed by AlphaFold2 for its design feasibility [7]. The topological analysis demonstrated that only 16 of structures in the above contained β-sheets, which indicates a pronounced bias toward helical topologies [7]. Moreover, there still exist steric clashes in generated structures and this requires energy minimization or relaxation [7]. For LigandMPNN method, the coverage of test set is narrow. As the authors report, there are only 317 small molecules, 74 nucleic acids and 83 metal species in the sets, hence it is insufficient to capture more complex spatial scenarios [8]. For ProDomino method, there still exist high failure rates and the limitation in optical dynamic range in low-scoring candidate sites. Initially, 50% of low-scoring hybrids failed in the screens of Cas12a-PDZ and this gives rise to substantial variability [3]. Besides, the residual activity in the close state only allowed a three-fold reduction rather than a state which is complete shut off and this is suboptimal where fully binary control is required in the light-controlled Cas12a-LOV2 experiments [3].

3.2. Outlook

To solve these issues, several strategies have emerged. For FoldingDiff's designability shrinkage, ProteinMPNN method can be combined with deep learning based secondary screening to address this issue. In the binder design scenarios, the success rates are increased nearly an order of magnitude and the native-scaffold sequence recovery are improved from 32.9% to 52.4% by using AF2 or RF structural reliability and interface scoring as filters [9].

For the small-molecule coverage limitations of LigandMPNN, uMol method can compensate this challenge by directly predicting protein small molecule complexes from sequence. In addition, this method can create or screen many more ligand protein complexes without requiring crystal structures when it combined with PoseBusters and energy-relaxation physical checks. For example, uMol achieved an SR of only 18% in blind inference, but when a pocket was specified, the SR rose up to 45%. Moreover, after apply applying pLDDT thresholds that strongly enriched for credible results, the SR reached about 72% when the pLDDT > 80 and rose up to about 80% when the pLDDT > 85. It also yields 336 high-confidence complexes for potential dataset expansion when the pLDDT > 85 [5].

PinMyMetal method is used to accurately predict transition-metal binding pockets aiming to increase the representation of metal-binding sites, which can allow bulk annotation of metal sites in the LigandMPNN predicted structures. For instance, this method was trained on 4,984 transition-metal binding sites across 2,778 structures and it has far exceeded the LigandMPNN benchmark that only has 83 metal sites [10].

Finally, the DIP-seq method can be used prior to ProDomino to identify the regions permissive to domain insertion, which can avoid blind insertion into the low tolerance sites. As a result, the high failure rate at low-scoring insertion sites in ProDomino method can be reduced. As illustration, DIP-seq quantitatively identified CTD hot zones with the condition of z-scores > 2, which indicates that insertion success rates reliably high across repeated experiments in the potassium channel protein Kir2.1. This shows it can effectively search a safe insertion band and restrict the ProDomino candidate sites to those high tolerance zones should markedly decrease failure rates [6].

4. Conclusion

In summary, this article has discussed three kinds of methods for function-controllable protein generation. They are FoldingDiff, LigandMPNN and ProDomino respectively. Although each method can encode functional features into the generative process in sdifferent ways, all of them still

face significant barrier such as designability shrinkage, insufficient topological diversity, narrow test-set coverage, and high failure rates at insertion sites. There are, however, some ways to make these problems less serious. For example, ProteinMPNN secondary screening, uMol with strict physical validation, PinMyMetal for transition-metal pocket annotation, and DIP-seq for mapping tolerant insertion zones. The large-model generation in protein is moving from just getting the right folding to making designs that are truly functional and can be controlled. With the creation of common benchmarks and strong experimental feedback loops, these changes are ready to make generative protein design useful in drug discovery, synthetic biology, and the making of new functional materials.

References

[1] Wu K E, Yang K, Van den Berg R, Alamdari S, Zou J Y, Lu A X, Amini A P. Protein structure generation via folding diffusion. Nature Communications, 2024, 15: 1059.

[2] Dauparas J, Lee G R, Pecoraro R., et al. atomic context-conditioned protein sequence design using LigandMPNN. Nature Methods, 2025, 22: 717 – 723.

[3] Wolf B, Shehu P, Brenker L, Von Bachmann A, Kroell A S, Southern N, Holderbach S, Eigenmann J, Aschenbrenner S, Mathony J, Niopek D. Rational engineering of allosteric protein switches by in silico prediction of domain insertion sites. Nature Methods, 2025, 15: 412.

[4] Butten Schoen M, Morris G M, Deane C M. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences (PoseBusters). Chemical Science, 2024, 15: 3130 – 3139.

[5] Bryant P, Kelkar A, Guljas A, Clementi C, Noé F. Structure prediction of protein-ligand complexes from sequence information with Umol. Nature Communications, 2024, 15: 4536.

[6] Jänes J, Beltrao P. Deep learning for protein structure prediction and design—progress and applications. Molecular Systems Biology, 2024, 20 (3): 162 – 169.

[7] Khakzad H, Igashov I, Schneuing A, Goverde C, Bronstein M, Correia B. A new age in protein design empowered by deep learning. Cell Systems, 2023, 14 (11): 925 – 946.

[8] Bennett N R, Coventry B, Goreshnik I, Huang B, Allen A, Vafeados D, Peng Y P, Dauparas J, Baek M, Stewart L, DiMaio F, De Munck S, Savvides S N, Baker D. Improving de novo protein binder design with deep learning. Nature Communications, 2023, 14: 2625.

[9] Zhang H, Zhong J, Gucwa M, Zhang Y, Ma H, Deng L, Mao L, Minor W, Wang N, Zheng H. PinMyMetal: a hybrid learning system to accurately model transition metal binding sites in macromolecules. Nature Communications, 2025, 16: 3043.

[10] Coyote-Maestas W, He Y, Myers C L, Schmidt D. Domain insertion permissibility-guided engineering of allostery in ion channels. Nature Communications, 2019, 10: 290.