https://doi.org/10.1038/s41587-024-02150-1

Spotlight on protein structure design

Check for updates

New methods for protein design speed up workflows, but issues of training data availability and method optimization remain.

Engineered proteins have been part of our daily lives for quite some time. For example, optimized enzymes are used in food processing for enhancing flavor or to increase shelf life and as ingredients in detergents to break down proteins or fats in stains, and most therapeutic antibodies have been engineered in some way to reduce immunogenicity and enhance specificity. Only recently, however, have we seen tremendous changes in the way the protein engineering process is approached. This month's Focus issue explores the latest developments in protein design, which are making the technology more accessible than before.

While initial machine learning attempts to exploit structure prediction for protein design focused on generating structures that fold using methods like hallucination or inpainting, the field has quickly moved on to other methods that are better suited to the challenge of designing proteins for a particular function. Structure prediction and protein engineering are quite different goals, but the success of AlphaFold2 (ref. 1) and RoseTTaFold[2] has provided new inputs to the field of protein design. Diffusion-based methods invert a protein structure prediction network and can be conditioned on design properties of interest for a particular function. While some functions like protein binding are relatively easy to design with these methods, other design targets, including enzymes and membrane proteins[3], remain more difficult. Designing functions that do not occur in nature poses even more challenges but also promises rich rewards in synthetic biotechnology and bioactive material production.

Traditionally, protein design has been approached from two different angles: optimization of an existing protein[4] or complete de novo design[5]. Interestingly, there are an increasing number of approaches being developed that blur the line between these two ends of the spectrum, with all machine learning methods incorporating knowledge about existing proteins. They differ in how they use this information, and they range from simple generative methods that are trained on only a single family of proteins to large language models that incorporate information from a broad range of protein structures and that are conditioned on a per use-case basis[6,7]. There are also some very recent models that combine different approaches and model sequence and structure concurrently.

Regardless of the details of the method, the field of computational protein engineering has benefitted tremendously from the open-source release of software and code enabling some level of democratization of research activity. But as there is a trend for models to become ever larger — machine learning models with billions of parameters are not uncommon — it remains to be seen whether this trajectory can continue or whether cutting-edge protein design research will become limited to big companies and a few well-funded labs[8].

Moreover, comparison of model performances and accurate benchmarking remain a challenge for the field. While some purely computational metrics like self-consistency have been proposed and demonstrated to be useful comparators, ultimate proof of design success lies in experimental validation, currently limited to labs with suitable facilities and expertise[9]. There is also a need for suitable training data. One contributing factor for the success of protein folding prediction methods has been the Protein Data Bank, which contains a wealth of solved protein structures that are publicly available. As more companies engage in protein engineering, they generate proprietary, specialized training sets that complement the publicly available data and are tailored to a specific design challenge at hand[10].

In addition to the quality of training data, another cornerstone of machine learning method success is the optimization of the model architecture itself. Ideally, the architecture incorporates prior knowledge about the problem either as part of the loss function or directly in the structure of the neural network. For protein structure prediction, one example of prior knowledge incorporated by state-of-the-art methods is translational and rotational invariance; that is, the fact that a protein structure remains the same no matter how it is positioned in space. The best way to incorporate such prior knowledge for protein engineering remains an open research question. Methods that incorporate physical information, such as the conformational folding energy, into a deep learning architecture might be promising for problems for which training data is sparse. Not only could this provide solutions to modeling static structure predictions, but this information could also help model protein dynamics.

Overall, the field has reached an inflection point where computational methods substantially speed up the process of protein design with increased success rates in comparison to previous engineering strategies based on energy functions or rational design. In practice, this means that experimentally testing a much smaller number of designs is often sufficient to achieve success, alleviating the bottleneck of wet lab validation. It is not surprising that computational design pipelines have already been picked up by the pharmaceutical industry, but it will be exciting to see the field extend far beyond therapeutics to the design of protein circuits, either in combination with bioelectronics or within cells as potential biosensors; development of new photosynthetic proteins; and development of biodegradable materials, carbon sequestration or enzymes to break down pollutants. As this Focus issue discusses, the options are limitless.

Published online: 15 February 2024

References

1. Jumper, J. et al. *Nature* https://doi.org/10.1038/s41586-021-03819-2 (2021).
2. Baek, M. et al. *Science* https://doi.org/10.1126/science.abj8754 (2021).
3. Li, H. et al. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-01987-2 (2024).
4. Notin, P., Rollins, N., Gal, Y., Sander, C. & Marks, D. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-024-02127-0 (2024).
5. Chu, A. E., Lu, T. & Huang, P.-S. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-024-02133-2 (2024).
6. Hsu, C., Fannjiang, C. & Listgarten, J. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-02115-w (2024).
7. Madani, A. & Ruffolo, J. A. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-024-02123-4 (2024).
8. Doerr, A. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-02111-0 (2024).
9. Chica, R. A. & Ferruz, N. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-02120-z (2024).
10. Nuzhna, L. & van Stekelenburg, T. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-01938-x (2023).