#### A NEW ERA FOR PROTEIN DESIGN

New deep-learned tool designs novel proteins with high accuracy

This new software from the Baker laboratory designs proteins that actually work in the wet lab. And you can use it to design your own proteins too, right online.

This was going to happen, and I expected the Baker lab to be the first group to report it. But honestly I didn’t expect it to happen so quickly:

Reverse an AlphaFold-like neural network to feed it 3D structures and obtain from it protein sequences that fold accordingly. This by itself didn’t turn out to work quite well, but it inspired further strategies for machine learning-based protein design. And eventually this tool, called ProteinMPNN, came out, with which scientists can now design proteins that fold (and hence work) as they need.

ColabFold and even web app versions of ProteinMPNN are already online for everybody to use.

# Protein structure and protein design

As I have covered in previous articles on AlphaFold and protein modeling (see an [index of them here](https://lucianosphere.medium.com/here-are-all-my-peer-reviewed-and-blog-articles-on-protein-modeling-casp-and-alphafold-2-d78f0a9feb61)), protein sequences dictate how a protein will acquire a 3D structure (the fold) which in turn dictates what functions it can exert, as well as its stability, solubility, etc. (For the biologists: I’m leaving aside the whole other universe of intrinsically disordered proteins.)

It is very often interesting to tackle the opposite problem: given a function that should be achieved by a given 3D structure (or given any other trait that one wants to optimize, such as stability), what protein sequence do we need (or what mutations on a starting sequence)?

This problem is in general coined *protein design*; it has several goal-specific sub-problems of which creating a whole protein from scratch is the hardest.

So far, while subproblems such as stabilizing existing proteins are increasingly addressed through machine learning, the problem of creating a whole new protein sequence from scratch has been treated mainly through physics-based methods. Without any doubt, the leader group in the domain is the Baker lab at the University of Washington in Seattle, which is actually running a whole Institute for Protein Design.

This group, also developer of protein modeling programs such as RoseTTAFold (less known than AlphaFold but apparently almost as accurate) quickly saw how the new machine learning technologies aimed at predicting protein structures could be reversed to predict which sequences would fold as desired. The problem seems trivial, but involves several computer engineering challenges and then the ultimate wall that protein design campaigns usually hit: synthesizing the predicted proteins experimentally and verifying that they truly fold as expected, and even better if they perform the expected function.

So far, the Baker lab’s best tool was the Rosetta toolbox, a multiverse of tools for protein structure prediction and design built upon a mainly physics-based model. Despite several stunning protein designs published in high-impact journals, the truth is that success rates are very low: only a small fraction of the Rosetta designs actually fold and work as expected.

# Machine learning for protein design

Now, the Baker lab created a totally new tool called ProteinMPNN that builds on machine learning to produce protein sequences from expected structures. Although many works already theorized this, ProteinMPNN is the first one proved through experimental means to actually produce protein sequences with a high chance of getting folded as expected. In other words, this means that when the experimental part of the group took the designed sequences produced by the program and tried to produce the encoded proteins in the wet lab, they actually got them; and moreover, when they solved their structures they matched the expected structures, in many cases also carrying the expected function.

As the name suggests, ProteinMPNN is built around a message passing neural network (MPNN). The core MPNN used in this work builds on previous work, even pre-AlphaFold2 !

#### [Generative Models for Graph-Based Protein Design](https://proceedings.neurips.cc/paper/2019/hash/f3a4ff4839c56a5f460c88cce3666a2b-Abstract.html" \t "_blank)

[Generative Models for Graph-Based Protein Design Part of Advances in Neural Information Processing Systems 32 (NeurIPS…](https://proceedings.neurips.cc/paper/2019/hash/f3a4ff4839c56a5f460c88cce3666a2b-Abstract.html" \t "_blank)

[proceedings.neurips.cc](https://proceedings.neurips.cc/paper/2019/hash/f3a4ff4839c56a5f460c88cce3666a2b-Abstract.html" \t "_blank)

The starting network consisted in 3 encoder and 3 decoder layers and 128 hidden dimensions, and predicted protein sequences in an autoregressive manner from N to C terminus using protein backbone geometric features built from CA positions (CA being the central carbon atom of an amino acid). The new work improved on this by incorporating also the positions of N, C and O backbone atoms plus a virtual CB atom, and also on improving how the network is propagated.

The ProteinMPNN network operates by passing distances between N, CA, C, O and virtual CB atoms through an encoder module to obtain graph nodes and edges. These features are then converted into amino acid probabilities at each site of the protein sequence by a decoder module that randomly samples amino acids from a set of all possible permutations. Finally, the highest probability can be casted into exact protein sequences to then try to produce these candidate proteins in the wet lab. (Often a pool of likely sequences are tested experimentally to maximize chances that one will work, and even before this there’s usually deep human expert inspection of the candidate designs -but this is out of the scope and focus of this article.)

Diagram

Description automatically generated

Figure by author based on usage of the software, open material, and own drawings.

Very importantly, while the original MPNN decoded the sequence from N to C terminus, ProteinMPNN performs this randomly and allows the user to pre-set (and fix) certain amino acids. This way, the protein sequence is built around the fixed parts, which would usually include regions that want to be fixed to achieve a function. For example an epitope if one wants to design a protein that will display it on its surface to work as a vaccine, or even a whole protein if one intends to design a protein that will bind to it.

# Main tests and applications

First, by training the ProteinMPNN model on thousands of high-resolution structures from the protein Data Bank the authors found that the extended geometric description was indeed helpful to better recover known sequences, performing substantially better than with CA positions only. Moreover, the fully trained model recovers sequences much better than the standard Rosetta-based approaches.

Next, by optimizing the range over which backbone geometry influences amino acid identity the authors concluded that performance saturated at “only” 32–48 neighbors. This means that the model is relative small, hence that it runs very fast. Indeed, as they report, ProteinMPNN runs over 200 times faster than their Rosetta protocol -besides producing better designs.

Last, the authors verified that running the designed sequences through AlphaFold 2 resulted in back-prediction of the designs -an independent indication that the sequence had a good chance of folding correctly.

# Applications

None of this survives the hype if the designed proteins don’t actually work., or at least they fold as intended. Well, as the preprint shows, a large fraction of the designed sequences are very soluble, have high expression levels, and crystalize well. So much, that the authors present cases where they rescued previously failed designs they had attempted with Rosetta.

The authors also showed that ProteinMPNN produces more realistic proteins than an alternative method based on protein sequence hallucination with AlphaFold 2. The proteins proposed by AlphaFold contained too many hydrophobic clusters, resulting in insolubility, while ProteinMPNN’s designs were largely more soluble -also stable, and in the cases for which structures were determined, also very close to the designs.

Moreover, ProteinMPNN’s proteins proved to actually fold as designed include monomers, cyclic homo-oligomers, tetrahedral nanoparticles, and target binding proteins, the latter essential to produce new kinds of vaccines, protein switches, and other proteins with biotechnological applications mediated by binding.

As I wrap up this article, a second preprint came out by the Baker lab that presents specific applications of ProteinMPNN to the design of a wide range of symmetric protein homo-oligomers given only a specification of the number of protein copies and the number of amino acids in the protein. Of course, proving experimentally that the proteins fold as intended.

Among highlights, the authors describe designs of giant rings with over 1500 amino acids, complex symmetries, and wide (10 nanometer) openings. These examples in particularly differ considerably from structures available in the Protein Data Bank, highlighting that the rich diversity of new protein structures that can be created is not limited to what’s already known. Overall, this work could pave the way for the design of more complex protein-based nanomachines such as nanopores for DNA sensing, nanomotors, antiviral nanoparticles, and more.

You’ll find the links to the two preprints in the readings suggested at the end.

# Closing notes and how to use ProteinMPNN yourself today

What the latest editions of the [Critical Assessment of Structure Prediction (CASP)](https://lucianosphere.medium.com/the-critical-assessment-of-structure-prediction-casp-over-a-quarter-century-tracking-the-state-bde0a92b3680) revealed is that machine learning models like AlphaFold can predict protein structures very well. Now a new field is opened by their reversal: create new proteins that fold as we want. In fact as the first author of the work tweeted, ProteinMPNN has become “the standard approach at the Institute of protein Design” due to “the high rate of experimental success and applicability to almost any protein sequence design problem”:

The tool is available as a “Quick Demo” notebook, but presumably more notebooks will come up soon:

#### [ProteinMPNN/quickdemo.ipynb at main · dauparas/ProteinMPNN](https://github.com/dauparas/ProteinMPNN/blob/main/colab_notebooks/quickdemo.ipynb" \t "_blank)

[Code for the ProteinMPNN paper. Contribute to dauparas/ProteinMPNN development by creating an account on GitHub.](https://github.com/dauparas/ProteinMPNN/blob/main/colab_notebooks/quickdemo.ipynb" \t "_blank)

[github.com](https://github.com/dauparas/ProteinMPNN/blob/main/colab_notebooks/quickdemo.ipynb" \t "_blank)

And this has been wrapped (work by Simon Duerr from

#### [EPFL Tech4Impact](https://medium.com/u/7dbe1c70684b?source=post_page-----41ae2a7d23d8--------------------------------)

) into a HuggingFace web app with which you can go now right away do a test:

#### [ProteinMPNN - a Hugging Face Space by simonduerr](https://huggingface.co/spaces/simonduerr/ProteinMPNN" \t "_blank)

[Discover amazing ML apps made by the community](https://huggingface.co/spaces/simonduerr/ProteinMPNN" \t "_blank)

[huggingface.co](https://huggingface.co/spaces/simonduerr/ProteinMPNN" \t "_blank)

Here’s for example an example run, with the amino acid probabilities and 10 proposed protein sequences -result obtained in less than 5 seconds:

Graphical user interface, text

Description automatically generated

Output from a quick test on trying to recover human ubiquitin.

# Readings and related resources

Preprint with the full presentation of the method, by Dauparas et al:

#### [Robust deep learning based protein sequence design using ProteinMPNN](https://www.biorxiv.org/content/10.1101/2022.06.03.494563v1" \t "_blank)

[While deep learning has revolutionized protein structure prediction, almost all experimentally characterized de novo…](https://www.biorxiv.org/content/10.1101/2022.06.03.494563v1" \t "_blank)

[www.biorxiv.org](https://www.biorxiv.org/content/10.1101/2022.06.03.494563v1" \t "_blank)

Preprint with applications to innovating protein assemblies, by Wicky et al:

#### [Hallucinating protein assemblies](https://www.biorxiv.org/content/10.1101/2022.06.09.493773v1" \t "_blank)

[Deep learning generative approaches provide an opportunity to broadly explore protein structure space beyond the…](https://www.biorxiv.org/content/10.1101/2022.06.09.493773v1" \t "_blank)

[www.biorxiv.org](https://www.biorxiv.org/content/10.1101/2022.06.09.493773v1" \t "_blank)

Another recent preprint, theoretical only, that also demonstrates a model that samples protein sequences and structures to accommodate a given motif -essentially a design task:

#### [Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem](https://arxiv.org/abs/2206.04119" \t "_blank)

[Construction of a scaffold structure that supports a desired motif, conferring protein function, shows promise for the…](https://arxiv.org/abs/2206.04119" \t "_blank)

[arxiv.org](https://arxiv.org/abs/2206.04119" \t "_blank)

(and check [this summary of the preprint in a Twitter thread by its first author](https://twitter.com/brianltrippe/status/1535063427894624256))

The Baker lab at the University of Washington:

#### [Main - Baker Lab](https://www.bakerlab.org/" \t "_blank)

[Welcome! We seek to understand the fundamental principles underlying protein structure and function, to encode these…](https://www.bakerlab.org/" \t "_blank)

[www.bakerlab.org](https://www.bakerlab.org/" \t "_blank)

## Related blog entries:

#### [What’s Up after AlphaFold on ML for Protein Structure Prediction?](https://towardsdatascience.com/whats-up-after-alphafold-on-ml-for-structural-biology-7bb9758925b8" \t "_blank)

[Will the AI-powered revolution in biology keep going? Can we expect a new breakthrough? What’s going on right now in…](https://towardsdatascience.com/whats-up-after-alphafold-on-ml-for-structural-biology-7bb9758925b8" \t "_blank)

[towardsdatascience.com](https://towardsdatascience.com/whats-up-after-alphafold-on-ml-for-structural-biology-7bb9758925b8" \t "_blank)

#### [Here are all my peer-reviewed and blog articles on protein modeling, CASP, and AlphaFold 2](https://lucianosphere.medium.com/here-are-all-my-peer-reviewed-and-blog-articles-on-protein-modeling-casp-and-alphafold-2-d78f0a9feb61" \t "_blank)

[I compiled here all my peer-reviewed articles (some papers, a couple of reviews, one opinion) and blog entries about…](https://lucianosphere.medium.com/here-are-all-my-peer-reviewed-and-blog-articles-on-protein-modeling-casp-and-alphafold-2-d78f0a9feb61" \t "_blank)

[lucianosphere.medium.com](https://lucianosphere.medium.com/here-are-all-my-peer-reviewed-and-blog-articles-on-protein-modeling-casp-and-alphafold-2-d78f0a9feb61" \t "_blank)