Better AI For Understanding Life on Earth: Predict First, Design Later

Yana Bromberg* Amarda Shehu†

Abstract

Generative AI is generating much enthusiasm on potentially advancing biological design in computational biology. In this paper we take a somewhat contrarian view, arguing that a broader and deeper understanding of existing biological sequences is essential before undertaking the design of novel ones. We draw attention, for instance, to current protein function prediction methods which currently face significant limitations due to incomplete data and inherent challenges in defining and measuring function. We propose a "blue sky" vision centered on both comprehensive and precise annotation of existing protein and DNA sequences, aiming to develop a more complete and precise understanding of biological function. By contrasting recent studies that leverage generative AI for biological design with the pressing need for enhanced data annotation, we underscore the importance of prioritizing robust predictive models over premature generative efforts. We advocate for a strategic shift toward thorough sequence annotation and predictive understanding, laying a solid foundation for future advances in biological design.

Keywords: AI-enabled biological design; bionformatics; generative AI; function prediction; annotation.

1 Introduction

Bioinformatics is experiencing a Generative AI (GenAI) fever. Even when severely restricting ourselves to macromolecules, it seems as if every other day new research is published on yet another deep neural network (DNN) expanding the footprint of GenAI over protein, DNA, and RNA sequences and even entire genomes. The gains seem nothing short of spectacular, including generating protein sequences with novel catalytic activities [1], novel antibody sequences with high expression yield and binding affinity [2], and even whole microbial genomes [3]. These capabilities are powered by foundation models – task-agnostic DNNs trained in a self-supervised manner to learn inherent data representations.

It is impossible not to be impressed by the sheer rate of and the diversity of publications (if not advances) in an emerging domain at the intersection of GenAI and biology — AI-enabled biological design. One would not be faulted for thinking that we now understand and control life. Case in point, the highlight in [4] places the model proposed one day earlier in [3] as having learned the language of DNA.

We respectfully disagree. We do not understand enough biology to claim that we can control it. Claims of control are premature at best, naive at worst. Through a somewhat controversial title, in this blue sky paper we argue that we should equally value and perhaps even prioritize *precisely annotating* the existing *diversity of sequences* rather than feverishly and single-mindedly focusing on *de-novo* synthesis. Our charge amounts to: *Fewer, potentially dubious bragging rights; more meaningful research.*

The reason for what we are suggesting can be summed up in one word: evolution. The shift in focus that we advocate recognizes the value of billions of years of evolution, this wonderful experiment that has generated a remarkable diversity of life forms characterized by intricate and highly regulated systems. While significant progress has been made in protein and nucleotide analysis using deep learning, the emphasis on generating novel sequences overlooks the vast, well-optimized results of natural evolution. We propose that understanding existing biological sequences should precede attempts to design new ones. Annotating sequences presents a significant challenge, but the potential rewards are substantial. Deeper insights into the complexities of biological systems can potentially identify new avenues for medical and industrial applications.

2 Data Before Models

It took the world's scientific research community nearly fifty years [5] and hundreds of millions of dollars to extract the protein 3D structure data present in the Protein Data Bank (PDB) [6] at the time of AlphaFold's appearance on the scene [7]. At its baseline, PDB is genius in its simplicity — for every protein amino acid sequence entry, there is a corresponding set of atom 3D coordinates. In "data-for-training" terms, this one-to-one mapping of language (sequence) to meaning (structure) was ideal.

Note that the biological diversity of the training sequences, necessary for the development of useful models, had not been a guaranteed or expected result of "normal" PDB data collection, where experimental work tends to favor large families of similar proteins. This diversity was

*Department of Computer Science, Department of Biology, Emory University, Atlanta, GA. Email: yana.bromberg@emory.edu

†Department of Computer Science, George Mason University, Fairfax, VA. Email: amarda@gmu.edu

externally imposed on the PDB [8] by the Structural Genomics Initiative [9] in an effort to create a more representative protein structure space. This effort, in part, was due to the desire to answer the question "Can we predict the protein structure from sequence?" — a question that has been thought about and addressed by the field since at least the first iteration of Critical Assessment of Structure Prediction (CASP) challenge in 1994 [10].

In fact, NNs have been championed as an answer to various aspects of protein structure prediction starting as early as 1993 [11] and have been consistently improved over time. Surely, deep learning advances have helped in the process! But would the DeepMind team been able to deliver such an impactful result in the absence of a well-defined question and/or lacking the PDB?

3 Data and Knowledge before Models

The road to AlphaFold was not a straight line. Scientific inquiry often reaches a tipping point through the convergence and integration of insights from multiple branches of knowledge. As AlQuraishi reflected on the performance of AlphaFold (v1.0) in 2019 at the biennial CASP competition [12]: "Their approach builds on two ideas developed in the academic community during the preceding decade: (i) the use of co-evolutionary analysis to map residue covariation in protein sequence to physical contact in protein structure, and (ii) the application of deep neural networks to robustly identify patterns in protein sequence and co-evolutionary couplings and convert them into contact maps." Co-variation, which identifies correlated mutations across protein sequences, provides insights into structural constraints driven by evolutionary pressures. This concept is grounded in a profound understanding of sequence evolution and the relationships between sequence, structure, and function. In other words, AlphaFold is not only a remarkable milestone but also a powerful testament to the decades of bioinformatics research that laid the essential groundwork of knowledge for its development.

The larger point we are making is that seeking to understand (and leverage) the well-optimized results of natural evolution keeps proving itself a worthy endeavor. The success of protein language models (PLMs), such as TAPE-BERT [13], Protein-BERT [14], ESM1b [15], ESM2 [16], Prottrans-BERT, Prottrans-Albert, and Prottrans-T5 [17], made possible by their ability to ingest millions of protein sequences, something that was not possible before with methods based on sequence alignment, is indeed further evidence of the benefit of "harnessing" the results of evolution (as in, for instance, feeding them directly to large models).

Yet, it is important to get things right. A growing argument in the scientific community is that PLMs implicitly learn structure due to their ability to ingest millions of protein sequences. The hypothesis is that this ability in turn enables capturing the selective pressures exerted on protein sequences throughout billions of years of evolution. In [18], the authors challenge this hypothesis. They "stress-test" PLMs on their ability to perform remote homology prediction, which requires structural knowledge to identify proteins with low sequence similarity. The results show that, while PLMs are better at this task than traditional sequence alignment methods, they still struggle in the "twilight zone" of very low sequence identity, exposing that they have not learned protein structure sufficiently. So, back to structure.

4 The Importance of Asking Questions: What Should we Know and Why?

Why do we want to know about protein structure? Beyond basic physics and biology, protein structure – the most-likely end result of the folding process – is important because it conveys certain aspects of protein function. Function, in turn, is relevant to drug development, understanding of disease, industrial advances, and eco-relevant engineering, among many other uses.

While it may come as a surprise to many AI researchers, scientists disagree about what exactly protein function is: Is it the interaction with other members of a biological pathway (e.g. binding of a metal)? Is it the purpose of such interaction (e.g. electron transfer)? Is it the resulting organismal phenotype (e.g. a live cell with plenty of energy)? This limited agreement on what function is means that only some definitions are accepted by some communities for some proteins [19, 20, 21, 22, 23, 24]. There is no one "functional PDB."

Why is it important to acknowledge this? Doing so affirms that we lack both the gold standard data relating a protein to its function (this protein does X) and the gold standard description of said function (what exactly is X?). Our limited understanding has implications for GenAI. Designing a new protein that performs a specific function requires a deep understanding of how that function is achieved by existing proteins. Without this knowledge, it is difficult to ensure that newly designed sequence will function as intended and will not disrupt existing biological systems.

We do not aim to add to the confusion, but it is important to acknowledge that we do not have answers to these basic questions. This is the first step that lays the foundation for what we are proposing; that is, our blue sky vision.

5 Blue Sky Idea(s)

Understanding biological function, even when restricted to macromolecules, is a *wicked problem*. While this term was first introduced by design theorist Horst Rittel and urban planner Melvin Webber in 1973 to describe challenges in social planning, we find it increasingly being used in the AI community and its sub-communities, including, for example, the growing "AI for science" sub-community. While its usage is often performative, we posit that indeed under-

standing and predicting biological function checks all the boxes: *no definitive problem statement* – indeed, function cannot be really defined in isolation of numerous interdependent factors (e.g. molecular interactions); *no clear path to a solution* – in the absence of a definitive answer to "what is function?" all we see are numerous (but limited) advances on related but isolated working definitions; *multiple stakeholders* with different viewpoints, interests, and definitions of success – check; *entwinement with other problems* – "solving" one aspect of function can create new problems or exacerbate existing ones in other biological system areas.

We posit key questions for the community:

- If we can not predict the function of a given protein sequence, is it safe to assume that we can design a new sequence with the desired functionality?
- How specific do we have to be in understanding and annotations to be able to design with precise control?
- What same things do we need to measure for a shared understanding and evaluation of our capabilities?
- What level of precision in the silicon leads to guaranteed viability in the petri dish?

It is worth instantiating these questions in hallmark wicked problems in bioinformatics. One such is *variant effect prediction*. It has been disheartening to see this problem reduced to overly-simplified (and in so doing, largely irrelevant) versions in the frenzy of claiming wins and flag poles for GenAI/PLMs. In [25] the authors remind the scientific community of what missense variant effect prediction is: the analysis of the impact of single amino acid substitutions resulting from single-nucleotide variants in genome coding regions. There is perhaps no more wicked word than "impact!" How do you define impact? How do you measure it? How do you annotate sequence data with impact?

These questions need to be formulated and answered precisely, as they affect the data, and through the data the models, and through the models the results. The latter can then either represent true advances, or fool us into thinking we are making progress, or, perhaps worse, obfuscate and confuse and stifle true advances and scientific research.

Specifically, work in [25] argues that the historical focus on high-impact and (human) pathogenic variants in experimental analyses used as training data has led to a skewed perception of variant effects. The authors explicitly ask whether the tools available today can *annotate* variants *accurately and comprehensively across related, but diverse categories* of effect: evolutionary fitness, pathogenicity, and functional change. They demonstrate that while traditional supervised methods are effective, they are constrained by biases in training data that disproportionately represent high-impact and pathogenic variants. In contrast, unsupervised PLMs perform comparably or even better than supervised methods in identifying functional and pathogenic variants, but further refinement and optimization are needed to establish a comprehensive understanding of variant effects and to enhance the precision of predictions, especially for less-studied organisms. The authors argue that achieving a "gold standard" predictor requires a clearer definition of variant effect in the first place, a move away from simplistic binary classifications and towards the development of larger, more balanced training sets that better represent the full spectrum of life and allow for inference of subtle or context-dependent impacts.

The blue sky vision we propose in this paper centers around *annotation*, with two key dimensions to it: *broader annotation*, as in expanding it to the diversity of data/organisms (that we do not yet have); and *precise annotation*, as in formulating categories (on what to annotate). This requires a whole-of-community approach and starts with asking the right questions, agreeing on definitions (what we are capturing through the annotations), and challenging ourselves to increasingly deeper and more specific categories of annotations (away from binary categories and deeper into the increasingly more precise and difficult instantiations of function, effect, impact, etc.). We hold that this approach is critical to truly advance biology *in silico* in all its beautifully wicked complexity. Before we rush to claim that we understand the rules of life (when we cannot even agree on fundamental properties), we need to proceed with caution (or in the words of philosophers, "epistemic humility"), critically reflecting on what is missing.

To ground this in yet another concrete setting: how do we know that AlphaFold has learned physics rather than memorized sequence patterns? If life has no other means of folding sequences than the ones we see (and research seems to point in this direction [26]), answering this question does not matter. Yet, if somewhere in the deep ocean there are sequences that are unlike anything we have seen at this point, then how do we know that AlphaFold can predict their structure? Could there be sequences out there that fold differently, that bind differently, that behave differently?

We need to lay a solid foundation for future advances in biological design, and we hold that this is only possible with a strategic shift toward comprehensive and precise sequence annotation and predictive understanding. Our vision is not *more data is better*; we are indeed proposing a deliberate approach for both diversity and precision. We underscore the importance of equally valuing and even perhaps prioritizing robust predictive models over premature generative efforts. This can be accomplished in many ways, from creating and sustaining specific communities, to being clear eyed as to what constitute true advances in peer reviewing, to valuing the things that matter in study sections and review panels. As we map out what nature has for us, we will be better positioned to design what nature has missed in its $\sim$4 billion year experiment.

Acknowledgments

This work is supported in part from NSF Grant No. 2310113 to AS and NSF Grant No. 2310114 to YB.

References

[1] A. Madani, B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, *et al.*, "Large language models generate functional protein sequences across diverse families," *Nature Biotechnology*, vol. 41, no. 8, pp. 1099–1106, 2023.

[2] N. Gruver, S. Stanton, N. Frey, T. G. Rudner, I. Hotzel, J. Lafrance-Vanasse, A. Rajpal, K. Cho, and A. G. Wilson, "Protein design with guided discrete diffusion," *Advances in neural information processing systems*, vol. 36, 2024.

[3] E. Nguyen, M. Poli, M. G. Durrant, B. Kang, D. Katrekar, *et al.*, "Sequence modeling and design from molecular to genome scale with evo," *Science*, vol. 386, no. 6723, p. eado9336, 2024.

[4] C. V. Theodoris, "Learning the language of DNA," *Science*, vol. 386, pp. 729–730, 2024.

[5] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. Meyer Jr, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, "The protein data bank: a computer-based archival file for macromolecular structures," *Journal of molecular biology*, vol. 112, no. 3, pp. 535–542, 1977.

[6] H. M. Berman, "The protein data bank: a historical perspective," *Acta Crystallographica Section A: Foundations of Crystallography*, vol. 64, no. 1, pp. 88–95, 2008.

[7] J. Jumper, R. Evans, *et al.*, "Highly accurate protein structure prediction with alphafold," *nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[8] R. Nair, J. Liu, T.-T. Soong, T. B. Acton, J. K. Everett, *et al.*, "Structural genomics is the largest contributor of novel structural leverage," *Journal of structural and functional genomics*, vol. 10, pp. 181–191, 2009.

[9] H. M. Berman, J. D. Westbrook, M. J. Gabanyi, W. Tao, *et al.*, "The protein structure initiative structural genomics knowledgebase," *Nucleic acids research*, vol. 37, no. suppl_1, pp. D365–D368, 2009.

[10] J. Moult, J. T. Pedersen, R. Judson, and K. Fidelis, "A large-scale experiment to assess protein structure prediction methods," 1995.

[11] B. Rost and C. Sander, "Improved prediction of protein secondary structure by use of sequence profiles and neural networks.," *Proceedings of the National Academy of Sciences*, vol. 90, no. 16, pp. 7558–7562, 1993.

[12] M. AlQuraishi, "Alphafold at casp13," *Bioinformatics*, vol. 35, pp. 4862–4865, 05 2019.

[13] R. Rao, N. Bhattacharya, N. Thomas, Y. Duan, P. Chen, J. Canny, P. Abbeel, and Y. Song, "Evaluating protein transfer learning with tape," *Advances in neural information processing systems*, vol. 32, 2019.

[14] N. Brandes, D. Ofer, Y. Peleg, *et al.*, "Proteinbert: a universal deep-learning model of protein sequence and function," *Bioinformatics*, p. 2102–2110, 2022.

[15] A. Rives, J. Meier, T. Sercu, *et al.*, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *Proc Natl Acad Sci USA*, vol. 118, no. 15, 2021.

[16] Z. Lin, H. Akin, *et al.*, "Language models of protein sequences at the scale of evolution enable accurate structure prediction," *bioRxiv*, 2022.

[17] A. Elnaggar, M. Heinzinger, C. Dallago, *et al.*, "Prottrans: Towards cracking the language of lifes code through self-supervised deep learning and high performance computing," *IEEE Trans Patern Anal Mach Intell*, vol. 44, no. 10, pp. 7112–7127, 2022.

[18] A. Kabir, A. Moldwin, Y. Bromberg, and A. Shehu, "In the twilight zone of protein sequence homology: Do protein language models learn protein structure?," *Bioinformatics Advances*, vol. 4, no. 1, p. vbae119, 2024.

[19] S. A. Aleksander, J. Balhoff, S. Carbon, J. M. Cherry, H. J. Drabkin, *et al.*, "The gene ontology knowledgebase in 2023," *Genetics*, vol. 224, no. 1, p. iyad031, 2023.

[20] P. D. Karp, R. Billington, R. Caspi, C. A. Fulcher, M. Latendresse, *et al.*, "The biocyc collection of microbial genomes and metabolic pathways," *Briefings in bioinformatics*, vol. 20, no. 4, pp. 1085–1093, 2019.

[21] I. U. of Biochemistry. Nomenclature Committee and I. U. of Biochemistry, *enzyme nomenclature, 1978: recommendations of the nomenclature Committee of the International Union of biochemistry on the nomenclature and classification of enzymes.* Academic Press, 1979.

[22] A. Chang, L. Jeske, S. Ulbrich, J. Hofmann, J. Koblitz, *et al.*, "Brenda, the elixir core data resource in 2021: new developments and updates," *Nucleic acids research*, vol. 49, no. D1, pp. D498–D508, 2021.

[23] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, *et al.*, "Kegg: Kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 27, no. 1, pp. 29–34, 1999.

[24] P. Bansal, A. Morgat, K. B. Axelsen, V. Muthukrishnan, E. Coudert, *et al.*, "Rhea, the reaction knowledgebase in 2022," *Nucleic acids research*, vol. 50, no. D1, pp. D693–D700, 2022.

[25] Y. Bromberg, A. Kabir, P. Ramakrishnan, and A. Shehu, "Variant effect prediction in the age of machine learning," *Cold Spring Harbor Perspectives in Biology*, p. a041467, 2024.

[26] Y. Bromberg, A. A. Aptekmann, Y. Mahlich, L. Cook, S. Senn, *et al.*, "Quantifying structural relationships of metal-binding sites suggests origins of biological electron transfer," *Science advances*, vol. 8, no. 2, p. eabj3984, 2022.