Session introduction: AI-driven Advances in Modeling of Protein Structure

Krzysztof Fidelis†

Protein Structure Prediction Center and Genome Center,
University of California, Davis,
Davis, CA 95616, USA
Email: kfidelis@ucdavis.edu

Sergei Grudinin

LJK CNRS, Université Grenoble-Alpes
Grenoble, 38000, France
Email: sergei.grudinin@univ-grenoble-alpes.fr

The last few years mark dramatic improvements in modeling of protein structure. Progress was initially due to breakthroughs in residue-residue contact prediction, first with global statistical models and later with deep learning. These advancements were then followed by an even broader application of the deep learning techniques to the protein structure modeling itself, first using Convolutional Neural Networks (CNNs) and then switching to Natural Language Processing (NLP), including Attention models, and to Geometric Deep Learning (GDL). The accuracy of protein structure models generated with current state-of-the-art methods rivals that of experimental structures, while models themselves are used to solve structures or to make them more accurate.

Looking at the near future of machine learning applications in structural biology, we ask the following questions: Which specific problems should we expect to be solved next? Which new methods will prove to be the most effective? Which actions are likely to stimulate further progress the most? In addressing these questions, we invite the 2022 PSB attendees to actively participate in session discussions.

The AI-driven Advances in Modeling of Protein Structure session includes five papers specifically dedicated to:
(1) Evaluating the significance of training data selection in machine learning.
(2) Geometric pattern transferability, from protein self-interactions to protein-ligand interactions.
(3) Supervised versus unsupervised sequence to contact learning, using attention models.
(4) Side chain packing using SE(3) transformers.
(5) Feature detection in electrostatic representations of ligand binding sites.

Keywords: Artificial intelligence; Machine learning; Deep learning; Natural language processing; Attention models; Graph convolutional networks; Geometric learning; Geometric vector perceptron; SE(3) transformers; Training data; Global statistical models; Protein structure modeling; Contact prediction; Side-chain modeling; Protein-ligand interactions; Ligand binding sites.

† Work partially supported by NIH grant R01 GM100482.
1. A short retrospect

It has been approximately 50 years since Christian Anfinsen posed the protein folding problem (Anfinsen, Haber et al. 1961) and Cyrus Levinthal formulated his famous paradox (Levinthal 1969). Ensuing, in addition to experimental studies, multiple theoretical and computational endeavors converged on understanding of macromolecular structure and function. These included a substantial investment in molecular mechanics, dynamics, and related simulations (Levitt and Warshel 1975, McCammon, Gelin et al. 1977, Van Gunsteren and Berendsen 1977, Brooks and Karplus 1983, Noguti and Gō 1983, Levitt, Sander et al. 1985, Abagyan and Mazur 1989, Mazur and Abagyan 1989). While of great theoretical interest (2013 Nobel prize in chemistry awarded to Karplus, Levitt, and Warshel), these methods could not reliably produce accurate protein structure models. For decades, the only successful approach in this area, comparative or homology modeling, had to rely on the knowledge of related, already known structures. To assess the effectiveness of methods across the protein structure modeling field and to stimulate progress, the Critical Assessment of Structure Prediction (CASP) experiments were launched in 1994 (Moult, Pedersen et al. 1995). Substantial advancements in the field were observed over the next two decades. These resulted in the development of multiple new approaches to modeling and to the increase in accuracy, notably for proteins without homologous structures available to draw upon (Das and Baker 2008, Zhang 2008). While protein size remained an important limitation, these developments significantly widened the range of modeling applications (Baker and Sali 2001, Moult 2008). But these advancements only foreshadowed further, even more significant progress. In 2014, the 11th edition of CASP showed how deep multiple sequence alignments (MSAs) could be used to dramatically increase model accuracy, whenever such data were available (Ovchinnikov, Kim et al. 2016). This result was possible through extracting covariation signals (coevolution) from MSAs, using global statistical models, such as direct-coupling analysis (Weigt, White et al. 2009, Marks, Colwell et al. 2011, Morcos, Pagnani et al. 2011), pseudolikelihood maximization (Balakrishnan, Kamisetty et al. 2011, Kamisetty, Ovchinnikov et al. 2013), or sparse inverse covariance estimation (Jones, Buchan et al. 2012), which successfully dealt with previous errors. CASP12 (2016), saw the first applications of deep learning (DL). First, treating a covariation signal as an image, followed by DL, i.e., a move from unsupervised statistical models to supervised DL, lead to further improvements in contact prediction (Wang, Sun et al. 2017). Second, DL architectures for 3D objects were designed and applied for the first time (Derevyanko, Grudinin et al. 2018). Just two years later (CASP13, 2018), these developments led to successful modeling of a wide range of proteins for which no template data were available. The feat was accomplished with the first version of the AlphaFold architecture (from a company DeepMind), which relied on the MSAs and the convolutional neural networks (CNNs) (Senior, Evans et al. 2020). Finally, in CASP14 (2020), the model accuracy on single protein domains reached a near-experimental level (again, with a DeepMind originating, AlphaFold2 architecture) (Jumper, Evans et al. 2021). This was made possible thanks to the direct processing of the MSAs (i.e., a raw-MSA method that did not rely on statistical models) (Ju, Zhu et al. 2021), the use of large meta-genomic databases and DL models, such as attention, brought from natural language processing (NLP) (Vaswani, Shazeer et al. 2017), rather than CNNs, and progress in deep geometric learning (Bronstein, Bruna et al. 2021). Of note, was the highly accurate estimation of per residue error (Jumper, Evans et al. 2021).
2. A brief outline of current research

In CASP14, the DeepMind team demonstrated that it was possible to predict highly accurate 3D models of proteins, with accuracy competitive with that of experimental structures (Jumper, Evans et al. 2021). Their architecture captures long-range dependencies between amino acid residues, which are transformed into structural constraints, while preserving symmetry, and properties of the 3D space.

These latest, but also previous accomplishments of the deep learning techniques in this area have stimulated the community to revisit protein sequence and structure representations. Advances in deep learning in the treatment of language (Vaswani, Shazeer et al. 2017) and 3D geometry (Bartok, Kondor et al. 2013, Cohen and Welling 2016, Gilmer, Schoenholz et al. 2017, Thomas, Smidt et al. 2018, Bronstein, Bruna et al. 2021) have resulted in new approaches for structural biology. These include language models trained on large numbers of sequences that can learn 3D contacts and distance information (AlQuraishi 2019, Bhattacharya, Thomas et al. 2020, Rao, Ovchinnikov et al. 2020, Rao, Liu et al. 2021); novel geometrical representations treating proteins as point clouds, molecular graphs, surfaces or Voronoi tessellations (Sverrisson, Feydy et al. 2020, Baek, DiMaio et al. 2021, Baldassarre, Menendez Hurtado et al. 2021, Hiranuma, Park et al. 2021, Igashov, Olechnovic et al. 2021, Igashov, Pavlichenko et al. 2021, Jing, Eismann et al. 2021); novel ways to learn coevolution signals (Baek, DiMaio et al. 2021, Ju, Zhu et al. 2021, Jumper, Evans et al. 2021); truly end-to-end architectures (Jumper, Evans et al. 2021, Kandathil, Greener et al. 2021), and more.

3. Future developments (complexes, ligand interactions, other molecules, dynamics, language models, geometry models, sequence design)

Development of protein complex prediction, guided by coevolution signals between complex partners, is already underway (Baek, DiMaio et al. 2021, Evans, O’Neill et al. 2021, Humphreys, Pei et al. 2021, Pozatti, Kundrotas et al. 2021). Beyond interactions, understanding of protein multiple states and flexibility is necessary to further knowledge of binding, enzymatic reactions, transport, and more. We have already seen DL techniques designed to reconstruct protein structural heterogeneity in Cryo-EM maps (Punjani and Fleet 2021, Rosenbaum, Garnelo et al. 2021, Zhong, Bepler et al. 2021). Predicting protein flexibility and its multiple states, starting from a single sequence, must be the next step.

Even a single mutation can have a dramatic effect on protein structure and function. Recent studies open the possibility that protein language models, pre-trained on millions of unlabeled protein sequences, can be fine-tuned with small amount of labeled data to predict effects of mutations (Madani, McCann et al. 2020, Biswas, Khimulya et al. 2021, Rives, Meier et al. 2021).

Also, end-to-end deep neural networks, designed to predict inter-residue distances or protein structures, could perhaps be inverted to guide sequence design with a specific structure or function goal in mind (Anishchenko, Chidyausiku et al. 2020, Norn, Wicky et al. 2021).

The current success of protein structure prediction is heavily grounded in the availability of large protein sequence datasets, in a relatively small fraction complemented by 3D protein structures. To
any comparable extent, data are not available for other types of macromolecules, like RNAs, and more geometry-based learning approaches need to be developed for these tasks.

4. **What is needed for further progress?**

To advance the deep learning approaches in structural bioinformatics, we need large, high-quality datasets, like those found in protein sequence and structure databases. We may also need additional collaborative platforms to facilitate scientific exchange. The large-scale, blind, method performance assessment experiments, such as CASP and other CASP-modeled platforms, including Critical Assessment of Protein Interactions (CAPRI) (Janin 2002), Critical Assessment of Genome Interpretation (CAGI) (Repo, Moult et al. 2012), and others, have already proved to be a motivating factor for the academic and later even industrial teams. But in the end, this is to a considerable degree an open question and we strongly invite your input.

5. **Overview of papers in this session**

5.1. **Evaluating significance of training data selection in machine learning**

*Derry, Carpenter, and Altman* examine performance of machine learning in three categories, (1) assessment of model accuracy, (2) design of protein sequence, and (3) catalytic residue prediction, when different types of experimental structural data are used in both evaluating and training of these techniques. Machine learning, and especially the deep learning methods, have demonstrated strong performance on these tasks, but they still very much depend on the data used in training. To assess this dependence, the authors use several datasets constructed from X-ray crystallography, nuclear magnetic resonance (NMR), and cryo-electron microscopy (cryo-EM) structural data available in the PDB. The study is timely, especially with the dramatic developments in cryo-EM and the resulting shift in structure data distribution. Performance is benchmarked using the GPV (Geometric Vector Perceptron) graph neural architecture (for accuracy and design) and a graph convolutional network (for function).

How well are the known biochemical and biophysical effects replicated in the trained models? How do the complex biases that affect structure determination influence the training outcome? The use of which training data produces the most reliable results? Can training be effectively optimized for a particular machine learning task? Does mixing data types improve performance? Does balancing the data with respect to the experimental data type improve performance? I their work, the authors, to a considerable degree, provide answers to these questions.

5.2. **Geometric pattern transferability**

*Koehl, Jagota, Erdmann-Pham, Fung, and Song* examine the geometric pattern transferability from protein self-interactions to protein-ligand interactions.

Exploring the intra-protein interactions as a possible training set substitute for protein-ligand interactions has a potential to alleviate the relative scarcity of the latter type in the PDB. The authors
use probabilistic models to characterize protein self-contacts and assess transferability with several statistical analyses. They then assess the results using an established protein-ligand docking dataset.

In their paper, the authors strive to provide answers to the following underlying questions: Do the amino acid chemical group pairs have similar geometric distributions in protein self-contacts and protein-ligand complexes? Which ligand contact geometries are not well represented in proteins? When evaluated using a protein-ligand docking protocol, to what degree is the overall performance acceptable?

5.3. Supervised versus unsupervised sequence to contact learning

Bhattacharya, Thomas, Rao, Dauparas, Koo, Baker, Song and Ovchinnikov investigate whether attention models and ProtBert-BFD type Transformers can meaningfully contribute to residue-residue contact prediction, and if the long-standing dependence on the co-evolution-based analysis using multiple sequence alignments (MSAs) can be reduced or eliminated. They also explore if the hierarchical structure within and across protein families can be a good source of signal for the Transformer attention models.

Can these methods effectively substitute for the established approach relying on the MSA-based training? Can these results lead to the development of useful protein representation models?

5.4. Side chain packing using SE(3) transformers

Jindal, Zhu, Chowdhury, Vajda, Padhorny and Kozakov use a 3D equivariant neural network architecture, specifically the SE(3)-transformers with a self-attention mechanism for 3D point cloud and graph data, to predict side chain conformations. The architecture they use adheres to equivariance constraints, to ensure that point cloud data are invariant to changes in the input pose. Their specific focus is on the protein-protein interfaces, critical in modeling of protein complexes.

Can this approach reduce or eliminate the need for combinatorial search in addressing the side chain rotamer selection problem? Is it robust enough for modeling of protein-protein interfaces? Can it become a useful part of a larger network architecture addressing protein structure modeling?

5.5. Feature selection in electrostatic representations of ligand binding sites

Quintana, Kong, He and Chen use voxel representation of electrostatic isopotentials and convolutional neural networks to classify electrostatic representations of ligand binding sites. They follow with class activation mapping to identify regions of electrostatic potential that are significant in these classifications. Finally, they argue that regions that drive classification are also likely to play a biochemical role in effecting binding specificity. Their DeepVASP-E algorithm is part of a series of methods they plan to develop within the Analytic Ensemble package, designed to discern biochemical mechanisms.
References


