HAL
open science

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

Krzysztof Fidelis, Sergei Grudinin

To cite this version:

Krzysztof Fidelis, Sergei Grudinin. Session introduction: AI-driven Advances in Modeling of Protein Structure. Pacific Symposium on Biocomputing 2022, Jan 2022, Fairmont Orchid - Hawaii, United States. hal-03874552

HAL Id: hal-03874552

https://hal.science/hal-03874552v1

Submitted on 28 Nov 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Biocomputing 2022 Downloaded from www.worldscientific.com
by 129.88.37.137 on 11/28/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

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.

© 2021 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

Biocomputing 2022 Downloaded from www.worldscientific.com
by 129.88.37.137 on 11/28/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

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 Leventhal 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).

Biocomputing 2022 Downloaded from www.worldscientific.com
by 129.88.37.137 on 11/28/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

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

Biocomputing 2022 Downloaded from www.worldscientific.com
by 129.88.37.137 on 11/28/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

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

Biocomputing 2022 Downloaded from www.worldscientific.com
by 129.88.37.137 on 11/28/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

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.

Biocomputing 2022 Downloaded from www.worldscientific.com
by 129.88.37.137 on 11/28/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

References

Abagyan, R. A. and A. K. Mazur (1989). "New Methodology for Computer-Aided Modelling of Biomolecular Structure and Dynamics 2. Local Deformations and Cycles." Journal of Biomolecular Structure and Dynamics **6**(4): 833--845.

AlQuraishi, M. (2019). "End-to-End Differentiable Learning of Protein Structure." Cell Syst **8**(4): 292-301 e293.

Anfinsen, C. B., E. Haber, M. Sela and F. H. White, Jr. (1961). "The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain." Proc Natl Acad Sci U S A **47**: 1309-1314.

Anishchenko, I., T. M. Chidyausiku, S. Ovchinnikov, S. J. Pellock and D. Baker (2020). "De novo protein design by deep network hallucination." bioRxiv.

Baek, M., F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read and D. Baker (2021). "Accurate prediction of protein structures and interactions using a 3-track network." Science: eabj8754.

Baker, D. and A. Sali (2001). "Protein structure prediction and structural genomics." Science **294**(5540): 93-96.

Balakrishnan, S., H. Kamisetty, J. G. Carbonell, S. I. Lee and C. J. Langmead (2011). "Learning generative models for protein fold families." Proteins **79**(4): 1061-1078.

Baldassarre, F., D. Menendez Hurtado, A. Elofsson and H. Azizpour (2021). "GraphQA: protein model quality assessment using graph convolutional networks." Bioinformatics **37**(3): 360-366.

Bartok, A. P., R. Kondor and G. Csanyi (2013). "On representing chemical environments." Phys. Rev. B **87**: 184115.

Bhattacharya, N., N. Thomas, R. Rao, J. Daupras, P. Koo, D. Baker, Y. S. Song and S. Ovchinnikov (2020). "Single Layers of Attention Suffice to Predict Protein Contacts." bioRxiv.

Biswas, S., G. Khimulya, E. C. Alley, K. M. Esvelt and G. M. Church (2021). "Low-N protein engineering with data-efficient deep learning." Nature Methods **18**(4): 389--396.

Bronstein, M. M., J. Bruna, T. Cohen and P. Velivcković (2021). "Geometric deep learning: Grids, groups, graphs, geodesics, and gauges." arXiv:2104.13478.

Brooks, B. and M. Karplus (1983). "Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor." Proceedings of the National Academy of Sciences **80**(21): 6571--6575.

Cohen, T. S. and M. Welling (2016). "Steerable CNNs." arXiv:1612.08498.

Das, R. and D. Baker (2008). "Macromolecular modeling with rosetta." Annu Rev Biochem **77**: 363-382.

Derevyanko, G., S. Grudinin, Y. Bengio and G. Lamoureux (2018). "Deep convolutional networks for quality assessment of protein folds." Bioinformatics **34**(23): 4046-4053.

Evans, R., M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A.

Biocomputing 2022 Downloaded from www.worldscientific.com
by 129.88.37.137 on 11/28/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper and D. Hassabis (2021). "Protein complex prediction with AlphaFold-Multimer." bioRxiv.

Gilmer, J., S. S. Schoenholz, P. F. Riley, O. Vinyals and G. E. Dahl (2017). Neural Message Passing for Quantum Chemistry. Proceedings of the 34th International Conference on Machine Learning. D. Precup and Y. W. Teh, PMLR. **70:** 1263--1272.

Hiranuma, N., H. Park, M. Baek, I. Anishchenko, J. Dauparas and D. Baker (2021). "Improved protein structure refinement guided by deep learning based accuracy estimation." Nature communications **12**(1): 1--11.

Humphreys, I. R., J. Pei, M. Baek, A. Krishnakumar, I. Anishchenko, S. Ovchinnikov, J. Zhang, T. J. Ness, S. Banjade, S. Bagde, V. G. Stancheva, X.-H. Li, K. Liu, Z. Zheng, D. J. Barrero, U. Roy, I. S. Fernández, B. Szakal, D. Branzei, E. C. Greene, S. Biggins, S. Keeney, E. A. Miller, J. C. Fromme, T. L. Hendrickson, Q. Cong and D. Baker (2021). "Structures of core eukaryotic protein complexes." bioRxiv.

Igashov, I., L. Olechnovic, M. Kadukova, C. Venclovas and S. Grudinin (2021). "VoroCNN: Deep convolutional neural network built on 3D Voronoi tessellation of protein structures." Bioinformatics **37**(16): 2332-2339.

Igashov, I., N. Pavlichenko and S. Grudinin (2021). "Spherical convolutions on molecular graphs for protein model quality assessment." Machine Learning: Science and Technology **2**(4): 1-12.

Janin, J. (2002). "Welcome to CAPRI: A Critical Assessment of PRedicted Interactions." Proteins-Structure Function and Genetics **47**(3): 257-257.

Jing, B., S. Eismann, P. Suriana, R. J. L. Townshend and R. Dror (2021). Learning from Protein Structure with Geometric Vector Perceptrons. International Conference on Learning Representations.

Jones, D. T., D. W. Buchan, D. Cozzetto and M. Pontil (2012). "PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments." Bioinformatics **28**(2): 184-190.

Ju, F., J. Zhu, B. Shao, L. Kong, T.-Y. Liu, W.-M. Zheng and D. Bu (2021). "CopulaNet: Learning residue co-evolution directly from multiple sequence alignment for protein structure prediction." Nature communications **12**(1): 1--9.

Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. \vZídek, A. Potapenko and others (2021). "Highly accurate protein structure prediction with AlphaFold." Nature: 1--11.

Kamisetty, H., S. Ovchinnikov and D. Baker (2013). "Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era." Proc Natl Acad Sci U S A **110**(39): 15674-15679.

Kandathil, S. M., J. G. Greener, A. M. Lau and D. T. Jones (2021). "Ultrafast end-to-end protein structure prediction enables high-throughput exploration of uncharacterised proteins." bioRxiv.

Levinthal, C. (1969). Mossbauer spectroscopy in biological systems. Proceedings of a meeting held at Allerton House. P. Debrunner, JCM Tsibris, and E. Munck, editors. University of Illinois Press, Urbana, IL.

Levitt, M., C. Sander and P. S. Stern (1985). "Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme." Journal of molecular biology **181**(3): 423--447.

Levitt, M. and A. Warshel (1975). "Computer simulation of protein folding." Nature **253**(5494): 694--698.

Biocomputing 2022 Downloaded from www.worldscientific.com
by 129.88.37.137 on 11/28/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

Madani, A., B. McCann, N. Naik, N. S. Keskar, N. Anand, R. R. Eguchi, P.-S. Huang and R. Socher (2020). "Progen: Language modeling for protein generation." arXiv preprint arXiv:2004.03497.

Marks, D. S., L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina and C. Sander (2011). "Protein 3D structure computed from evolutionary sequence variation." PLoS One **6**(12): e28766.

Mazur, A. K. and R. A. Abagyan (1989). "New methodology for computer-aided modelling of biomolecular structure and dynamics 1. Non-cyclic structures." Journal of Biomolecular Structure and Dynamics **6**(4): 815--832.

McCammon, J. A., B. R. Gelin and M. Karplus (1977). "Dynamics of folded proteins." Nature **267**(5612): 585--590.

Morcos, F., A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa and M. Weigt (2011). "Direct-coupling analysis of residue coevolution captures native contacts across many protein families." Proc Natl Acad Sci U S A **108**(49): 1293-1301.

Moult, J. (2008). "Comparative modeling in structural genomics." Structure **16**(1): 14-16.

Moult, J., J. T. Pedersen, R. Judson and K. Fidelis (1995). "A large-scale experiment to assess protein structure prediction methods." Proteins **23**(3): ii-v.

Noguti, T. and N. Gō (1983). "Dynamics of native globular proteins in terms of dihedral angles." Journal of the Physical Society of Japan **52**(9): 3283--3288.

Norn, C., B. I. Wicky, D. Juergens, S. Liu, D. Kim, D. Tischer, B. Koepnick, I. Anishchenko, D. Baker and S. Ovchinnikov (2021). "Protein sequence design by conformational landscape optimization." Proceedings of the National Academy of Sciences **118**(11).

Ovchinnikov, S., D. E. Kim, R. Y. Wang, Y. Liu, F. DiMaio and D. Baker (2016). "Improved de novo structure prediction in CASP11 by incorporating coevolution information into Rosetta." Proteins **84 Suppl 1**: 67-75.

Pozatti, G., P. Kundrotas and A. Elofsson (2021). "Improved protein docking by predicted interface residues." bioRxiv.

Punjani, A. and D. J. Fleet (2021). "3D Flexible Refinement: Structure and Motion of Flexible Proteins from Cryo-EM." bioRxiv.

Rao, R., J. Liu, R. Verkuil, J. Meier, J. F. Canny, P. Abbeel, T. Sercu and A. Rives (2021). "MSA transformer." bioRxiv.

Rao, R., S. Ovchinnikov, J. Meier, A. Rives and T. Sercu (2020). "Transformer protein language models are unsupervised structure learners." bioRxiv.

Repo, S., J. Moult, S. E. Brenner and C. Participants (2012). "CAGI: The Critical Assessment of Genome Interpretation, a community experiment to evaluate phenotype prediction." Journal of Medical Genetics **49**: S29-S29.

Rives, A., J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma and others (2021). "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences." Proceedings of the National Academy of Sciences **118**(15).

Rosenbaum, D., M. Garnelo, M. Zielinski, C. Beattie, E. Clancy, A. Huber, P. Kohli, A. W. Senior, J. Jumper, C. Doersch and others (2021). "Inferring a Continuous Distribution of Atom Coordinates from Cryo-EM Images using VAEs." arXiv preprint arXiv:2106.14108.

Senior, A. W., R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Zidek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu and D. Hassabis (2020). "Improved protein structure prediction using potentials from deep learning." Nature **577**(7792): 706-710.

Sverrisson, F., J. Feydy, B. Correia and M. Bronstein (2020). "Fast end-to-end learning on protein surfaces." bioRxiv.
Thomas, N., T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff and P. Riley (2018). "Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds." arXiv:1802.08219.
Van Gunsteren, W. and H. J. Berendsen (1977). "Algorithms for macromolecular dynamics and constraint dynamics." Molecular Physics **34**(5): 1311--1327.
Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin (2017). Attention Is All You Need. NeurIPS Proceedings.
Wang, S., S. Sun, Z. Li, R. Zhang and J. Xu (2017). "Accurate de novo prediction of protein contact map by ultra-deep learning model." PLoS computational biology **13**(1): e1005324.
Weigt, M., R. A. White, H. Szurmant, J. A. Hoch and T. Hwa (2009). "Identification of direct residue contacts in protein-protein interaction by message passing." Proc Natl Acad Sci U S A **106**(1): 67-72.
Zhang, Y. (2008). "I-TASSER server for protein 3D structure prediction." BMC Bioinformatics **9**(1): 40.
Zhong, E. D., T. Bepler, B. Berger and J. H. Davis (2021). "CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks." Nature Methods **18**(2): 176--185.

Biocomputing 2022 Downloaded from www.worldscientific.com
by 129.88.37.137 on 11/28/22. Re-use and distribution is strictly not permitted, except for Open Access articles.