Dean&Francis

ISSN 2959-409X

A Review of the Research Progress of Simplified Models in Protein Structure Prediction

Ang Li[1]

[1] *College of Agronomy and Biotechnology, The Southwest University, Chongqing, China*

* Corresponding author: al396s@missouristate.edu

Abstract:

Protein is the main executor of life activities. Proteins form specific three-dimensional conformations through the folding of amino acid sequences. The functions of proteins are decided by their three-dimensional structure. However, the use of experimental approaches to determine protein structures has numerous limitations. Protein folding methods for prediction are constantly evolving as a result of advances in computational biology and artificial intelligence. Simplified models are frequently employed to lower computational complexity due to the complexity of the all-atomic model and the unpredictable nature of the chemical reaction. This paper discusses some core issues in protein folding prediction, reviews relevant theories such as Levinthal's paradox, Energy Tunnel Theory, and Metadynamics, and summarizes the progress of research methods of simplified models in protein folding prediction. It was also explored how to improve the algorithm-based model. This article's goal is to review the current status of studies on simplified models for the identification of protein structures from the perspectives of biophysical and mathematical biology.

Keywords: Protein Folding; Simplified Models; Coarse-Grained Models; Machine Learning; Energy Landscape

1 Introduction

Proteins are the main executors of life activities. Simply put, the proteins' 3-D framework, which is created via folding, dictates their function. But once the folding process goes wrong, the consequences can be extremely serious. The misfolding of proteins is directly related to various diseases. The mechanism of these diseases largely stems from the failure of proteins to form the correct three-dimensional structure, resulting in the loss of normal function and toxicity.

However, there are many limitations to determining protein structure through experimental methods. For example, experimental methods often only display the final static structure and cannot demonstrate the dynamic folding path from amino acid sequence to natural conformation.

ISSN 2959-409X

Computational models are useful in this situation. With the advancement of techniques for forecasting folds of proteins using intelligent machines and biological computation are continuously evolving. Due to the complexity of all-atom models and the uncertainties surrounding energy functions, simplified models are widely used to reduce computational complexity.

As the prototype of simplified models, toy models reveal the essence of folding through abstract physical interactions. Toy Models include HP model, AB model, BLN model, and Tube model. The Hydrophobic Polar Model, proposed by Ken Dill in 1985, simplifies 20 amino acids into two categories: hydrophobic (H) and polar (P). It simulates folding through two-dimensional or three-dimensional lattice point self-avoidance walking, and only calculates the contact energy of non-bonded H residues. It is the first minimalist model that explicitly states that hydrophobic interactions dominate folding [1]. The AB model was proposed at the same time as HP, which uses continuous spatial simulation (non-lattice) of A (hydrophobic) and B (hydrophilic) residues to describe interactions through bond angles and Lennard Jones potentials and can capture the compact conformation of globulin [2]. The BLN model has added neutral residues (N), distinguishing helices, folds, and turns through dihedral potential energy, and for the first time, analyzes the differences in folding pathways of topologically similar proteins [3]. The Tube model uses C α atoms as flexible tube axes and simulates side chain stacking through pipe diameter and curvature constraints [4]. These models reveal the essential laws of protein folding through extreme simplification, especially the HP model.

However, using toy models to predict the precise arrangement of real proteins is not feasible, since they disregard side chain features and polarity interactions. Therefore, they need to be combined with coarsening to approximate the real system. A crucial methodological foundation for the coarse-grained model is provided by the simplified concept of toy models, such the HP model, which transforms hydrophobic interactions into calculable energy values. However, the coarse-grained model significantly improves its authenticity through dynamic mapping rules and multi-scale potential energy functions. The core idea of the coarse-grained model is to selectively preserve key interactions and merge redundant degrees of freedom, such as abstracting multiple atoms or functional groups into a single "coarse-grained" structure, to capture the core mechanisms of protein folding and function while reducing computational costs. Compared to toy models, coarse-grained models retain more physical details; Compared to the all-atom model, its computational efficiency is improved by 1 to 3 orders of magnitude, making it a core tool for simulating long-chain proteins or dynamic processes, and connecting theoretical exploration with practical applications [5].

However, the existing research focuses on the algorithm optimization or application of a single model, lacking a systematic combining of the evolutionary logic of "toy model coarse-grained model", especially ignoring the methodological enlightenment of simplified models on subsequent models. This article aims to trace the development process from toy models such as the HP model to modern coarse-grained models, with a focus on analyzing the core simplification ideas, algorithm evolution, and combination with experiments of the HP model. It reveals the inherent laws of the simplified model from physical abstraction to data fusion, providing insights for design.

2 The core issues of protein folding prediction

The fundamentals of anticipating protein folding are to decipher the mapping relationship between sequence, structure function. Its core problem can be divided into two dimensions: static structure prediction and dynamic path analysis, which combined provide a comprehensive grasp of the mechanism of protein folding.

2.1 Static Structure Prediction

Using the amino acid sequence to determine the natural 3-dimensional framework of proteins, that is, the stable conformation with the lowest energy or kinetically achievable, is the aim of static structure prediction. This structure serves as the foundation for protein function analysis. The core goal is to determine the natural three-dimensional structure of proteins, which is the lowest energy conformation that is thermodynamically stable or kinetically achievable, starting from the amino acid sequence. Protein function analysis is based on this. Its essence is to search for the global optimal solution in a high-dimensional conformational space. However, the conformational space's exponential complexity and the energy function's accuracy in recognizing natural states are its two main obstacles.

2.1.1 Levinthal's paradox

The Levinthal paradox is a classic theory in protein folding research that reveals the contradiction between folding speed and conformational spatial complexity. It was proposed by American biophysicist Cyrus Levinthal in 1968, and its core is a profound questioning of how proteins can quickly fold to their natural state. It is still a key starting point for understanding folding mechanisms to this day. On a theoretical level, the Levinthal paradox suggests that if proteins search for natural structures by randomly trying all possible conformations, the time required will far exceed the age of the universe. However, in reality, protein folding is usually completed within milliseconds to seconds [6]. Protein folding is not a random search, as this contradiction shows, but rather the existence of some effective guiding mechanism. Therefore, the key to static prediction is to narrow down the search space through simplification or guidance.

2.1.2 Energy Funnel Theory

To address this contradiction, the academic community has proposed the energy funnel hypothesis. The energy funnel hypothesis is a classic theory that explains how proteins fold from disordered polypeptide chains into specific three-dimensional structures in milliseconds, proposed by scholars such as Peter Wolynes in the 1990s. The fundamental idea is that proteins' energy domains are funnel-shaped rather than entirely random, with the natural state at the bottom, and the energy gradually decreases as the conformation approaches the natural state [7]. The folding process is guided by physical interactions and does not require traversing all conformations. Among them, the vertical axis of the funnel represents the conformational independence of energy, which is lowest in the normal state and largest in the unfolded form; The horizontal axis represents the entropy of the conformation, with the top unfolded having the highest entropy, corresponding to a large number of loose conformations, and the bottom natural state having the lowest entropy, corresponding to a single compact structure. The ramp mechanism refers to a funnel ramp where energy decreases and entropy decreases as the conformation approaches the natural state, but the gain in enthalpy exceeds the loss of entropy, resulting in a decrease in net free energy.

This theory suggests that during the folding process, proteins move along the "slope" of the funnel towards lower energy states, continuously eliminating high-energy erroneous conformations, and ultimately rapidly converging to their natural structure. This explains why folding can be so efficient because it is an energy-driven directional process.

2.2 Dynamic Path Analysis

2.2.1 Dynamic simulation of simplified models

Simplifying models by abstracting physical interactions reduces the complexity of dynamic simulations and becomes the main force for analyzing folding paths. The HP model, through lattice dynamics simulation, revealed for the first time the hydrophobic-driven pathway [8]. Short sequence simulations showed that folding begins with the formation of the first H-H contact, followed by the gradual aggregation of adjacent H residues and the exposure of P residues to the solvent, ultimately forming a compact conformation. Although the time scale of this process cannot directly correspond to real time, it is clear that hydrophobic interactions are the core driving force for path guidance. The dynamic modification of bond and dihedral angles in the folding of 9-residue chains is enhanced by the continuous spatial simulation of the AB model. In the initial stage, a loose conformation is formed through hydrophobic collapse, and then structural optimization is achieved through bond angle rotation [9]. The entire path exhibits a two-stage characteristic of "fast collapse slow optimization". The coarse-grained model reduces atomic degrees of freedom by 1-2 orders of magnitude while retaining key interactions, significantly extending simulation time [10].

2.2.2 Metadynamics

Metadynamics is an improved sampling technique designed to effectively explore the free energy landscape (FES) of complex systems while overcoming the time scale constraints of conventional molecular dynamics simulations. The core principle is to use selected collective variables (CVs) during the simulation process. In space, periodic deposition of Gaussian potential barriers forms repulsive bias potentials that accumulate over time. This bias potential will remember the regions that the system has already explored, forcing it to escape from local energy minima and cross energy barriers to explore new conformational spaces [11]. The cumulative bias potential progressively approaches the negative value of the target system's free energy as the simulation duration increases. Therefore, the free energy surface can be directly reconstructed through the evolution of the bias potential, revealing key information such as energy barriers, intermediate states, and stable conformations of the system. To address the potential energy oscillation issues in standard element dynamics, the improved Well Tempered Meta dynamically reduces the height of the Gaussian barrier, allowing the system to sample more smoothly at effective temperatures and further enhancing the convergence of the free energy surface [12]. In protein research, metadynamics can capture rare events during the folding process, such as the conformational transition from alpha helix to beta fold. Meanwhile, by selecting appropriate CVs, the mechanism of toxic intermediate state formation of disease-related proteins can also be revealed, providing key insights for understanding protein dynamic function and pathological mechanisms.

3 Evolution of Research Methods

3.1 Traditional Physical Models

The traditional physical model is based on the basic laws of molecular motion and explores the conformational changes and energy landscape of proteins through mathematical modeling and numerical simulation. Its core is to describe the folding process through computable physical interactions, providing a theoretical framework for simplifying models and modern methods.

3.1.1 Molecular Dynamics Simulation

Molecular Dynamics (MD) is a deterministic method for simulating molecular motion by solving Newtonian mechanical equations. Its core is to transform the motion of atoms or residues into the temporal evolution of position and velocity. The fundamental idea is to discretize Newton's formulas for motion in order to monitor the position of every particle in the system over time.

$$X(t+\Delta t) = X(t) + \Delta t\text{¡¤}v(t) + \Delta t^2\text{¡¤}\nabla V(X) \qquad (1)$$

V (t) is the atom's velocity at this instant, while X (t) is the atom's position at a specific time t. Additionally, the potential energy produced by the model that depicts atom-to-atom interaction is V (X) [6].

To introduce temperature effects, Langevin dynamics maintains the system temperature at the target value by adding random forces and friction terms to the motion equation, balancing physical realism and computational feasibility. The first protein application of MD can be traced back to the simulation of bovine trypsin inhibitors in 1977, but in the early days, it was limited by computing power and could only simulate nanosecond-level short processes. The breakthrough of distributed computing has greatly expanded its time scale. The benefit of MD is its capacity to offer dynamic details at the atomic level, including the minutiae of hydrogen bond creation and bond angle rotation [13]. However, the disadvantage is the limited time scale, which still poses challenges for long-chain proteins or slow processes.

3.1.2 Monte Carlo method

Random sampling is the foundation of the Monte Carlo (MC) method, a statistical mechanics technique, which explores the energy landscape of a system by constructing a Markov chain. It does not rely on deterministic dynamics and is more suitable for global optimization. Its core is the Metropolis Monte Carlo (MMC) algorithm, which works by randomly generating new conformations and accepting or rejecting them based on probability. The lower energy conformations are automatically accepted, while the high-energy conformations are accepted with a certain probability to avoid getting stuck in local minima. Unlike MD's "continuous motion", MC's "random jumping" is easier to cross energy barriers and is suitable for searching for global energy minima. For example, in the HP model, MC efficiently selects low-energy conformations by randomly changing the direction of lattice chains; In simulated annealing, MC combines temperature scheduling to accept more high-energy conformations at high temperatures to explore a wide range, and focuses on local optimization at low temperatures, successfully generating the natural state of Met enkephalin [14]. But the disadvantage of MC is the lack of time information, which makes it impossible to directly simulate the dynamic path.

3.1.3 Energy Function and Force Field

The energy function and force field are the engines of traditional physics models, used to quantify the interactions between molecules, and their accuracy directly determines the reliability of simulation results. The energy function is usually decomposed into chemical bonding and non-bonding interactions. The refinement of the force field refers to the optimization of parameters by fitting experimental data to the full atomic or force field with coarse particles. The core challenge of the energy function is to balance "simplification" and "precision". Oversimplification may lose key interactions, while over-complexity can increase computational costs. The evolution trend of the force field is to combine experimental data with data-driven methods to enhance its applicability to complex systems [15].

3.2 Optimization Algorithm

3.2.1 Ant Colony Algorithm

Ant Colony Optimization (ACO) is a heuristic improvement algorithm inspired by ant foraging behavior, which simulates the pheromone transmission mechanism to achieve global search and exhibits unique advantages in lattice problems such as HP models [16]. The core principle is that a group of artificial ants construct protein chain conformations by moving on lattice points, leaving pheromones on the path after each step of movement, which is negatively correlated with conformational energy; Subsequently, ants often select routes with high pheromone concentrations, and pheromones gradually dissipate to steer clear of local optima before coming together to form the ideal shape.

In the HP model, the adaptability of ACO is reflected in the discreteness of grid point movement. Ants start from the first residue of the sequence and extend their chain in the direction of grid points at each step. The legitimacy of their movement is constrained by self-avoidance, and the update of pheromones is linked to the quantity of H-H connections. The more contacts there are, the greater the increment of pheromones. ACO performs better than the genetic algorithm in 2D HP models for core sequences with dense hydrophobic residues [17], as its pheromone mechanism can better guide the aggregation of hydrophobic residues; But in long sequences, the efficiency is lower than that of Replica Exchange, mainly because the evaporation rate of pheromones is difficult to balance exploration and utilization.

The value of ACO lies in its swarm intelligence characteristics, which can avoid a single path from getting stuck in local minima through distributed search. However, its performance is highly dependent on parameters, and its application in continuous space models is limited. It is more commonly used as an auxiliary algorithm for lattice models.

3.2.2 Genetic Algorithm

By simulating the processes of selection, crossover, and mutation, genetic algorithms (GA), which are grounded on the concepts of genetics and natural selection, accomplish conformational optimization. They are a classic method for handling high-dimensional conformational spaces. The core steps in protein structure prediction.

Including encoding, fitness function, selection, crossover, and mutation. Firstly, transform the protein conformation into a chromosome, such as encoding the direction of the chain using lattice direction sequences in the HP model [18]. Secondly, the quality is measured by conformational energy, with lower energy indicating higher fitness ; Thirdly, select a high fitness conformation as the parent through roulette wheel or tournament selection [19]; Fourth, exchange partial fragments of the parent chromosome to generate offspring conformation; Finally, randomly changing a certain locus on the chromosome introduces new conformational diversity.

The genetic algorithm has been successfully applied in both HP model and the coarse-grained model. In the 2D HP model, its global minimum search efficiency for 20 residue sequences is 10-30 times that of standard Monte Carlo, as crossover operations can quickly combine high-quality segments of the parent generation. In the BLN model, the genetic algorithm combined with dihedral angle encoding successfully captured the low-energy conformation of spiral folding transition. But its limitation lies in the design of crossover and mutation operations: the high-dimensional dependence of protein chains may lead to the generation of invalid conformations through crossover, and repair mechanisms need to be combined to improve efficiency [20].

The strength of the genetic algorithm's worldwide search capability is its benefit. which can handle discrete or continuous conformational spaces in simplified models, and is easy to parallelize, serving as a bridge between simplified models and modern evolutionary algorithms.

The combination of simplifying models and optimizing algorithms essentially involves reducing the dimensionality of problems through model simplification and improving search efficiency through algorithm optimization. The HP model provides a standardized testing platform for algorithms, while ant colony algorithm, genetic algorithm, etc. use different heuristic strategies to break the complexity of conformational space, jointly promoting the transition of static structure prediction from theoretical exploration to practical application.

4 Discussion

4.1 Limitations of Current Research

Despite significant progress in simplifying models and optimizing algorithms for protein structure prediction, there still exists an irreconcilable contradiction between accuracy, efficiency, and universality.

The extreme simplification model only retains hydrophobic interactions and cannot capture the fine-tuning of structures by weak interactions such as hydrogen bonding and electrostatic interactions. Although the coarse-grained model increases the types of interactions, parameterization relies on empirical fitting, resulting in a sharp drop in accuracy during cross-system migration.

Even efficient algorithms still face the curse of dimensionality in the conformational search of long sequences or multi-domain proteins. Although the crossover or mutation operations of genetic algorithms and ant colony algorithms can introduce diversity, they may generate invalid conformations in continuous space models, requiring complex repair mechanisms and ultimately reducing efficiency [21].

Full atomic simulation is limited by microsecond-level time scales, and the dynamic results of simplified models are difficult to correlate to real time. Most dynamic simulations focus on a single path, making it difficult to explain the widespread phenomenon of multi-path folding in living organisms [22].

Modern hybrid models heavily rely on homologous sequence data, resulting in a sharp decline in prediction accuracy for orphan proteins. Although the simplified model does not rely on homologous information, it cannot distinguish sequence-specific interactions due to the insufficient universality of the potential energy function.

4.2 Improvements

The intervention of machine learning provides a new paradigm for protein structure prediction to break through traditional limitations. Its core is to learn the implicit association between protein sequences and structures through data-driven learning, while collaborating with physical models to achieve a closed loop of “data fitting physical constraints efficient search”. In recent years, machine learning has significantly expanded its value in basic research and practical applications by improving model accuracy, accelerating conformational search, and predicting dynamic paths.

4.2.1 Model Improvement

Data-driven potential function optimization: The potential function of traditional simplified models relies on empirical parameters and is difficult to adapt to sequence diversity. Machine learning achieves dynamic adjustment of the potential energy function by mining massive, structured data.

The CGSchNet model utilizes graph neural networks (GNNs) to learn all atom simulation data and represents the coarse-grained potential energy function as a nonlinear connection between position and distance between residues, which can automatically capture the interaction changes caused by mutations [23]. In a mixed force field, the combination of AlphaFold and coarse-grained models predicts residue distance and direction maps through deep learning, providing “prior constraints” for the physical potential energy function [24].

4.2.2 Algorithm Acceleration

Traditional optimization algorithms are prone to getting stuck in local optima in high-dimensional conformational spaces, while machine learning accelerates the search through a dimensionality reduction guidance strategy.

Variational autoencoder (VAE) can compress high-dimensional conformational space into low dimensional latent space and use simple algorithms such as gradient descent to find the optimal solution in the latent space before mapping it back to the original space.

5 Conclusion

Computational models are the core tools for analyzing protein folding mechanisms, evolving from toy models to coarse-grained models. The toy model simplifies abstraction to reveal core folding driving forces such as hydrophobic interaction, laying a methodological foundation for subsequent research. The coarse-grained model inherits the simplified approach and, by balancing physical details with computational efficiency, becomes the key to connecting theory with application.

Static prediction addresses the challenges of high-dimensional conformational space through algorithms such as ACO and GA, while dynamic analysis breaks through the limitations of time scale by means of meta-dynamics, etc. However, current research still has contradictions among accuracy, efficiency and universality.

Machine learning offers a new path to break through limitations. Data-driven optimization of potential energy functions, algorithmic level VAE dimensionality reduction to accelerate search, RNN learning of folded paths, etc.

In the future, it is necessary to deepen the integration of physical models and data-driven approaches to simplify the physical constraint design of machine learning guided by model logic, enhance the universality of data methods, and promote research from single structures to path prediction, moving towards systematic analysis of folding mechanisms, functional regulation, and disease pathology, serving protein engineering and drug design.

Machine learning methods still face issues such as data bias, physical interpretability, and dynamic generalization. In the future, it is necessary to integrate multiple sources of data to alleviate data bias. First, developing physically interpretable AI; Second, combining quantum computing to accelerate high-dimensional dynamic simulation, achieving a leap from static prediction to dynamic control. The deep integration of machine learning with simplified models and physical methods is reshaping the paradigm of protein structure prediction: shifting from rule-based simplification to data-driven intelligent simplification, and from passive search to active guidance. This fusion not only improves prediction accuracy and efficiency but also promotes innovation throughout the entire chain from basic research to industrial applications, providing unprecedented tools for solving the sequence structure function problem.

References

1. K. A. Dill. Theory for the folding and stability of globular proteins. Biochemistry, 24(6), 1501–1509. (1985). https://doi.org/10.1021/bi00327a032

2. F. H. Stillinger, T. Head-Gordon, & C. L. Hirshfeld. Toy model for protein folding. Physical review E, 48(2), 1469. (1993)

3. J. Kim, & T. Keyes. Inherent structure analysis of protein folding. The journal of physical chemistry. B, 111(10), 2647–2657. (2007). https://doi.org/10.1021/jp0665776

4. J. R. Banavar, M. Cieplak, & A. Maritan. Lattice tube model of proteins. Physical review letters, 93(23), 238101. (2004). https://doi.org/10.1103/PhysRevLett.93.238101

5. P. Banerjee, R. Lipowsky, & M. Santer. Coarse-Grained Molecular Model for the Glycosylphosphatidylinositol Anchor

with and without Protein. Journal of chemical theory and computation, 16(6), 3889–3903. (2020). https://doi.org/10.1021/acs.jctc.0c00056

6. M. Lougher, M. Lücken, T. Machon, M. Malcomson, & A. Marsden. COMPUTATIONAL MODELLING OF PROTEIN FOLDING.University Of Warwick. (2010)

7. P. G. Wolynes. Evolution, energy landscapes and the paradoxes of protein folding. Biochimie, 119, 218–230. (2015). https://doi.org/10.1016/j.biochi.2014.12.007

8. S. Moreno-Hernández, & M. Levitt. Comparative modeling and protein-like features of hydrophobic-polar models on a two-dimensional lattice. Proteins, 80(6), 1683–1693. (2012). https://doi.org/10.1002/prot.24067

9. Z. Li, & H. A. Scheraga. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. Proceedings of the National Academy of Sciences of the United States of America, 84(19), 6611–6615. (1987). https://doi.org/10.1073/pnas.84.19.6611

10. N. Singh, & W. Li. Recent Advances in Coarse-Grained Models for Biomolecules and Their Applications. International journal of molecular sciences, 20(15), 3774. (2019). https://doi.org/10.3390/ijms20153774

11. D. Ray, & M. Parrinello. Kinetics from Metadynamics: Principles, Applications, and Outlook. Journal of chemical theory and computation, 19(17), 5649–5670. (2023). https://doi.org/10.1021/acs.jctc.3c00660

12. M. Nava. Implementing dimer metadynamics using gromacs. Journal of computational chemistry, 39(25), 2126–2132. (2018). https://doi.org/10.1002/jcc.25386

13. D. Macuglia, B. Roux, & G. Ciccotti. The emergence of protein dynamics simulations: how computational statistical mechanics met biochemistry. The European Physical Journal H, 47(1), 13. (2022)

14. M. Roucairol, & T. Cazenave. Solving the hp model with nested monte carlo search. arXiv preprint arXiv:2301.09533. (2023)

15. G. A. Papoian (Ed.). Coarse-grained modeling of biomolecules. CRC Press. (2017)

16. M. D. Toksari. Ant colony optimization for finding the global minimum. Applied Mathematics and computation, 176(1), 308-316. (2006)

17. A. G. Citrolo, & G. Mauri. A local landscape mapping method for protein structure prediction in the HP model. Natural Computing, 13(3), 309-319. (2014)

18. S. P. Dubey, N. G. Kini, S. Balaji, & M. S. Kumar. A comparative study on single and multiple point crossovers in a genetic algorithm for coarse protein modeling. Critical Reviews™ in Biomedical Engineering, 46(2). (2018)

19. R. Matoušek. Genetic algorithm and advanced tournament selection concept. In Nature Inspired Cooperative Strategies for Optimization (NICSO 2008) (pp. 189-196). Berlin, Heidelberg: Springer Berlin Heidelberg. (2009)

20. F. L. Custódio, H. J. Barbosa, & L. E. Dardenne. Investigation of the three-dimensional lattice HP protein folding model using a genetic algorithm. Genetics and Molecular Biology, 27, 611-615. (2004)

21. M. K. Islam, M. Chetty, & M. Murshed. Conflict resolution based global search operators for long protein structures prediction. In International Conference on Neural Information Processing (pp. 636-645). Berlin, Heidelberg: Springer Berlin Heidelberg. (2011, November)

22. E. J. Guinn, B. Jagannathan, & S. Marqusee. Single-molecule chemo-mechanical unfolding reveals multiple transition state barriers in a small single-domain protein. Nature communications, 6(1), 6861. (2015)

23. N. E. Charron, K. Bonneau, A. S. Pasos-Trejo, A. Guljas, Y. Chen, F. Musil, J. Venturin, D. Gusew, I. Zaporozhets, A. Krämer, C. Templeton, A. Kelkar, A. E. P. Durumeric, S. Olsson, A. Pérez, M. Majewski,B. E. Husic, A. Patel, G. De Fabritiis, F. Noé, C. Clementi. Navigating protein landscapes with a machine-learned transferable coarse-grained model. Nature chemistry, 17(8), 1284–1292. (2025). https://doi.org/10.1038/s41557-025-01874-0

24. A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, & D. Hassabis. Improved protein structure prediction using potentials from deep learning. Nature, 577(7792), 706–710. (2020) https://doi.org/10.1038/s41586-019-1923-7