Machine and Deep Learning Innovations for Protein Structure Quality Assessment

Loubna Terra[1,*,†], Fouzia Benchikha[1,†] and Mohamed Hachem Kermani[2]

[1]Laboratory LIRE, Abdelhamid Mehri University Constantine 2 , Constantine , Algeria
[2]Laboratory LIRE, National Polytechnic School, Constantine , Algeria

Abstract

The importance of protein structures in biomedical research, especially in the drug discovery and design process, cannot be overlooked. The accuracy of these structures is crucial to ensure the success of research endeavors. However, experimental determination of protein structures is expensive and time-consuming, and computational predictions are not flawless. Therefore, assessing the quality of protein models has become a vital step in filtering the most reliable options before further exploration. To meet this need, various structural bioinformatics labs have developed methods for Evaluating Model Quality (EMQ). Applying machine learning (ML) to EMQ has emerged as one of the most effective approaches, as evidenced by the results of the CASP challenge, which is widely recognized within the scientific community. This article offers a systematic analysis of the leading ML-based EMQ methods developed in recent years. We categorize these methods based on the ML technology used and examine their relevance from a methodological perspective. We also introduce the fundamentals of EMQ. Overall, this article aims to serve as a starting point for exploring current research on protein quality evaluation while discussing future prospects in this rapidly evolving field.

Keywords

protein structure prediction, model quality assessment, machine learning (ML), deep learning (DL), CASP, EMQ.

1. Introduction

The fascinating process through which amino acids fold into three-dimensional protein structures is a natural wonder that plays a critical role in the myriad of functions executed by proteins within living organisms. Delving into the exact structures of proteins is essential for the advancement of molecular biology, biochemistry, and pharmacology, offering deep insights into the molecular mechanisms of life and fostering innovation in drug development, disease treatment, and the emerging field of synthetic biology. Traditionally, the determination of protein structures relied heavily on experimental methods such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy (cryo-EM) [1]. While these techniques have provided invaluable data, they are often constrained by high costs, technical challenges, and inherent limitations, such as the difficulty in crystallizing certain proteins or the extensive time requirements for data collection and analysis.

The 13th International Conference On Research in Computing at Feminine 2024 (RIF 2024)
*Corresponding author.
†These authors contributed equally.
loubna.terra@univ-constantine2.dz (L. Terra); fouzia.benchikha@univ-constantine2.dz (F. Benchikha); hachem.kermani@enp-constantine.dz (M. H. Kermani)

The emergence of computational methods for predicting protein structure (PSP) marks a paradigm shift, offering the promise of accelerating the pace of discovery while circumventing the limitations associated with traditional experimental approaches. Over recent decades, the field of PSP has witnessed substantial advancements, evolving from basic homology modeling techniques to sophisticated machine learning algorithms capable of predicting structures from amino acid sequences with remarkable accuracy. The integration of deep learning technologies, exemplified by the development of DeepMind's AlphaFold2 [2], represents a monumental leap in our ability to predict protein structures with near-experimental accuracy across a wide range of proteins. This breakthrough has set new benchmarks in the Critical Assessment of Structure Prediction (CASP) competitions, highlighting a significant stride forward in the realm of computational biology.

Concurrently, the importance of model quality assessment (QA) has been increasingly recognized, as it is essential for determining the reliability of predicted protein structures. QA methods enable the discernment of the most plausible models from a plethora of predictions, offering a measure of confidence in the models utilized for further biological interpretation or drug design. The evolution of QA methodologies has mirrored the advancements in PSP, with a notable shift towards the application of machine learning and deep learning techniques for a more nuanced analysis and interpretation of structural data.

This article endeavors to synthesize and compare various significant works that showcase the ongoing evolution and current status of PSP and QA methodologies. Each piece of work discussed represents a distinct contribution to the overarching effort to accurately predict and evaluate protein structures. By providing a comprehensive summary of these key ideas and methodologies, the article aims to offer a panoramic view of the advancements and challenges within the PSP and QA fields. It highlights the transformative impact of machine learning and deep learning technologies on our capabilities to predict and evaluate protein structures, paving the way for groundbreaking discoveries and applications in biology and medicine. As we continue to refine these computational tools, their integration into the broader ecosystem of structural biology promises to unlock new horizons in our understanding and utilization of the proteome.

2. Background

2.1. Machine learning and deep learning

Machine learning is a field of study within artificial intelligence (AI) focused on designing, analyzing, developing, and implementing methods that allow a machine (broadly defined) to evolve through a data-driven process rather than traditional deterministic algorithms. Machine learning approaches can be broadly classified into four types: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning [3]. Supervised learning acquires knowledge from training data with labeled responses. The learning process iteratively and automatically adjusts the internal parameters of the prediction model, aiming to minimize prediction errors. Most model quality assessment (MQA) methods are based on supervised machine learning algorithms.

Deep learning is a newer research domain within machine learning, introduced with the goal

of bringing ML closer to its ultimate objective: artificial intelligence. It involves algorithms inspired by the structure and function of the brain [4]. Deep learning encompasses a set of machine learning algorithms attempting to learn multiple levels of representation in order to model complex relationships between data. It has the capability to extract features from raw data through multiple layers of processing, consisting of multiple linear and nonlinear transformations, and to learn about these features gradually across each layer with minimal human intervention [3].

Machine learning/deep learning (ML/DL) algorithms and their use in MQA methods will be discussed in Section 3.

2.2. Protein structure prediction

Protein structure prediction, crucial for understanding biological functions, remains a key challenge in structural bioinformatics. This task involves inferring the three-dimensional structure of a protein from its amino acid sequence, with three main approaches distinguished: homology modeling [5] ,fold recognition[5] and ab initio prediction [4]. Recent advancements include the use of residue contact prediction, enriched by co-evolutionary analysis from multiple sequence alignments (MSA), significantly improving model accuracy [5].

The advent of deep learning, especially with algorithms such as DeepMind's AlphaFold [2], MULTICOM and RaptorX Contact, has marked a significant breakthrough, enabling accurate prediction of complex protein structures [5]. These methods leverage deep sequential and structural features to predict inter-residue distances and spatial configurations in a global context, setting new accuracy standards in the field.

2.3. Critical Assessment of Structure Prediction (CASP)

The Critical Assessment of Protein Structure Prediction (CASP), is a global competition aimed at evaluating protein structure prediction methods and fostering progress in the field[6]. Since its inception in 1994, CASP has played a crucial role in evaluating and advancing methods for predicting protein structures. Beginning with CASP7, the focus has extended to model quality evaluation (MQA) methods, which assess both global and local quality of protein structures submitted by prediction servers [5]. The integration of deep learning techniques, especially from CASP13, has significantly improved prediction accuracies, exemplified by methods like AlphaFold and RaptorX Contact [2]. These advancements challenge MQA methods to keep pace with the continuously improving model quality[5]. Prominent MQA methods in recent evaluations include FaeNNz, ModFOLD7, ProQ3D, and several others [5].

2.4. Model Quality Assessment Metrics

MQA methods are crucial for selecting the most accurate protein structure models from predictions, thus supporting biomedical research, particularly in drug discovery.

The evaluation metrics used to judge the accuracy of protein structure predictions include the Global Distance Test Total Score (GDT-TS)[5], Template Modeling (TM) score [7], local-Distance Difference Test (lDDT) score [8]and RMSD (Root Mean Square Deviation)[5]. These metrics assess the similarity between the predicted model and a reference experimental structure, focusing

on the ability to overlay sets of residues and measure the accuracy of inter-residue contacts.also, allow a comprehensive and nuanced evaluation of predicted model quality, contributing to the continuous improvement of protein structure prediction methods and the effectiveness of ML and DL-based MQA methods in CASP challenges.

3. EMQ methods based on ML and DL

This section compares several Model Quality Assessment (MQA) applications selected for their high popularity, immediate availability, and performance in CASP. Most of these methods are based on artificial neural networks (CNNs, GNNs).

Table 1 shows the details of these ML and DL-based MQA methods.

3.1. Method based on 3DCNN

The advanced approach using three-dimensional convolutional neural networks (3DCNN) represents a significant innovation in assessing protein structure quality, focusing on the detailed analysis of local quality to predict the overall model quality [9]. Using CASP dataset collections, this method enhances prediction accuracy through careful feature selection and optimized network topologies. It demonstrates the effectiveness of deep learning in protein structure evaluation, promising substantial advancements in structural bioinformatics [9].

3.2. TopQA

TopQA introduces an innovative method for protein structure quality assessment based on topology and employing machine learning. By leveraging a unique topological representation and applying a CNN to predict the GDT-TS score, TopQA surpasses traditional methods in accuracy, as evidenced by a correlation of 0.41 on CASP12 [10]. Developed from data in the CASP10 and CASP11 competitions, this approach optimizes the use of structural features for model training. TopQA, accessible via GitHub, signifies a progression in protein model evaluation by emphasizing their topological structure, opening new research avenues in structural bioinformatics [10].

3.3. SynthQA

SynthQA represents a breakthrough in protein model quality evaluation, utilizing a hierarchical architecture based on machine learning to analyze multi-scale features, from energetic scores to protein topology[11]. This method enhances evaluation accuracy over traditional approaches by analyzing and generating new features for optimized model training[11].

3.4. DeepUMQA

DeepUMQA is a cutting-edge method for evaluating protein structure quality, using ultra-rapid shape recognition (USR) and deep learning to effectively combine multi-scale features [12]. It stands out by surpassing well-established methods through its ability to detail structural

Table 1
Comparison of different EMQ methods

Name	Year and ref	Approach	Train,Test and Valid Data	Features	Metrics	Objective
3DCNN	2019, [9]	Deep convolutional neural networks (3D-CNN)	CASP7-10, CASP11, CASP12	Using the spatial characteristics of protein structures	GDT-TS	Predicting the quality of predicted protein structures using 3D convolutional neural networks
TopQA	2020, [10]	Topology-based machine learning	CASP10, CASP11, CASP12	Topological analysis of predicted protein structures	GDT-TS	Assess the quality of predicted protein structures by analyzing their topology
SynthQA	2021, [11]	Hierarchical machine learning	CASP10, CASP12, CASP14	Integration of hierarchical protein structure features	GDT-TS, lDDT and other customized metrics	Hierarchical assessment of the quality of protein structures predicted from amino acid sequences
DeepUMQA	2022, [12]	Ultra-fast pattern recognition based on deep learning	CASP13, CASP14, CAMEO	Use of geometric features for rapid recognition of protein structures	AUC, ROC, ASE	Accelerate quality assessment of predicted protein structures by focusing on essential geometric features
EnQA	2023, [13]	3D-equivariant graphical neural networks	CASP14, CAMEO, AlphaFold-train, AlphaFoldtest	Integration of structural features acquired from the state-of-the-art tertiary structure prediction method-AlphaFold2	GDT-TS, lDDT, and other custom metrics	Evaluate the quality of protein structural models, taking into account the rotation and translation of 3D objects

information of residues, proven by superior performance on the CASP13, CASP14, and CAMEO datasets [12].

3.5. EnQA

EnQA utilizes an innovative 3D equivariant graph neural network to assess protein model quality, leveraging advanced features from AlphaFold2 for accurate and transformation-insensitive evaluation [13]. This method exceeds the performance of traditional approaches and AlphaFold2 in quality assessment, illustrating its potential to transform protein structure evaluation in structural bioinformatics [13].

4. Discussion

Contemporary protein model quality assessment techniques like TopQA, SynthQA, DeepUMQA, and EnQA face significant challenges. These methods typically depend heavily on specific databases, which may not accurately represent the diversity of protein structures, potentially leading to biased outcomes. Additionally, they require substantial computational resources, restricting their use in environments with limited capabilities. Although these tools incorporate advanced deep learning and hierarchical architectures to enhance their evaluations, they often struggle to apply their findings beyond the initial training datasets. Consequently, models generally perform well on familiar data but fail to replicate this success on new, unseen datasets. This lack of robustness underscores the urgent need for innovation in structural bioinformatics to develop methods that are adaptive, less data-dependent, and more efficient, thus enhancing their reliability and practical utility across varied scenarios.

5. Conclusion

The evaluation of protein structures using ML and DL has demonstrated progress but also presents challenges, particularly in terms of generalization and data dependency. The limitations of current methods such as EMQ, observed during CASP competitions, underscore the need to explore more sophisticated and adaptive deep learning architectures. By adopting convolutional , residual or graph neural networks, we can expect improvements in prediction accuracy and an enhanced ability to process complex protein structures. The future of structural bioinformatics will heavily depend on our ability to integrate these advanced technologies, thereby ensuring significant advances in understanding biological functions and developing new therapies.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT for certain translation tasks. After using this tool, the authors reviewed and edited the content as needed, and assume full responsibility for the published material.

References

[1] J. Lengyel, E. Hnath, M. Storms, T. Wohlfarth, Towards an integrative structural biology approach: combining Cryo-TEM, X-ray crystallography, and NMR, Journal of Structural and Functional Genomics 15 (2014) 117–124. URL: http://link.springer.com/10.1007/s10969-014-9179-9. doi:10.1007/s10969-014-9179-9.

[2] 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 (2020) 706–710. URL: https://www.nature.com/articles/s41586-019-1923-7. doi:10.1038/s41586-019-1923-7.

[3] D. S. Sara, T. Loubna, T. Zineeddine, Approche Deep Learning basée sur le réseau de neurones résiduel pour la prédiction des séquences d'ADN virales. (2020).

[4] C. Hardin, T. V. Pogorelov, Z. Luthey-Schulten, Ab initio protein structure prediction (2002).

[5] J. Chen, S. W. I. Siu, Machine Learning Approaches for Quality Assessment of Protein Structures, Biomolecules 10 (2020) 626. URL: https://www.mdpi.com/2218-273X/10/4/626. doi:10.3390/biom10040626.

[6] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, A. Tramontano, Critical assessment of methods of protein structure prediction (CASP) — round x, Proteins: Structure, Function, and Bioinformatics 82 (2014) 1–6. URL: https://onlinelibrary.wiley.com/doi/10.1002/prot.24452. doi:10.1002/prot.24452.

[7] J. Xu, Y. Zhang, How significant is a protein structure similarity with TM-score = 0.5?, Bioinformatics 26 (2010) 889–895. URL: https://academic.oup.com/bioinformatics/article/26/7/889/213219. doi:10.1093/bioinformatics/btq066.

[8] V. Mariani, M. Biasini, A. Barbato, T. Schwede, lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests, Bioinformatics 29 (2013) 2722–2728. URL: https://academic.oup.com/bioinformatics/article/29/21/2722/195896. doi:10.1093/bioinformatics/btt473.

[9] R. Sato, T. Ishida, Protein model accuracy estimation based on local structure quality assessment using 3D convolutional neural network, PLOS ONE 14 (2019) e0221347. URL: https://dx.plos.org/10.1371/journal.pone.0221347. doi:10.1371/journal.pone.0221347.

[10] J. Smith, M. Conover, N. Stephenson, J. Eickholt, TopQA: a topological representation for single-model protein quality assessment with machine learning (2020).

[11] M. Korovnik, K. Hippe, J. Hou, D. Si, K. Kishaba, R. Cao, Synthqa - Hierarchical Machine Learning-Based Protein Quality Assessment, 2021. URL: http://biorxiv.org/lookup/doi/10.1101/2021.01.28.428710. doi:10.1101/2021.01.28.428710.

[12] S.-S. Guo, J. Liu, X.-G. Zhou, G.-J. Zhang, DeepUMQA: ultrafast shape recognition-based protein model quality assessment using deep learning, Bioinformatics 38 (2022) 1895–1903. URL: https://academic.oup.com/bioinformatics/article/38/7/1895/6520805. doi:10.1093/bioinformatics/btac056.

[13] C. Chen, X. Chen, A. Morehead, T. Wu, J. Cheng, 3D-equivariant graph neural networks for protein model quality assessment, Bioinformatics 39 (2023) btad030. URL: https://academic.oup.com/bioinformatics/article/doi/10.1093/bioinformatics/btad030/6986970.