

Advances in Computational Intelligence-Based Methods of Structure and Function Prediction of Proteins

Jian Zhang *  and Jingjing Qian

School of Computer and Information Technology, Xinyang Normal University, Xinyang 464000, China; 2023211107@xynu.edu.cn

* Correspondence: jianzhang@xynu.edu.cn; Tel.: +86-0376-6390715

Proteins serve as the building blocks of life and play essential roles in almost every cellular process. Protein sequence refers to the linear arrangement of amino acids, which are connected by peptide bonds. The protein sequence provides the blueprint for the corresponding secondary and three-dimensional (3D) structures [1,2]. As is well known, the protein's sequence determines its structure. However, given various environmental conditions, post-translational modifications, and interactions with other ligands, the same protein sequence may form different 3D structures [3]. Experimental structure determination and function analysis techniques are usually time-consuming and labor-intensive [4], which can benefit from computational intelligence-based models and approaches [5]. Computational-based methods have significantly contributed to protein structure prediction by offering a powerful suite of tools and algorithms that can accurately model and simulate the 3D conformations of proteins. These methods are invaluable in bridging the gap between the protein's primary sequence and the complex, functional architecture of a protein's tertiary shape. CASP (Critical Assessment of Structure Prediction) is a community that invites participants to provide their models to identify unknown protein structures [6]. This competition is held every two years. The latest CASP15 in 2023 attracted about 100 groups of researchers and received over 53,000 models on 127 modeling targets [7]. These methods of prediction of protein structures include: (i) template-based modeling from known protein structures; (ii) ab initio modeling via sophisticated energy functions or conformational search techniques; and (iii) artificial intelligence-aided methods that employ machine learning or deep learning algorithms. Table 1 collects the representative predictors for each category mentioned above. We selected three popular predictors for each type of protein structure prediction method. The citations are used as a direct way to quantify the impact of these resources within the community [8]. The citation counts were collected from Google Scholar (<https://scholar.google.com/>) on 16 August 2024.

Specifically, template-based modeling methods attract the most attention among bioscientists. These methods require high-resolution, accurate protein structure/template databases [9]. This type of method tries to find the most similar templates to model the structure of the target, unknown protein. The template-based modeling methods are regarded as the most successful and widely used approaches [9,10]. However, they may fail when proper orphan proteins or templates are not available. Then, it usually relies on pre-designed physical principles and statistical potentials to predict the structure only from the protein primary sequence. This type of method belongs to ab initio modeling. Recent years have witnessed the rapid development of artificial intelligence (AI) techniques. With the development of computing power, especially the rise of GPU computing, GPU-accelerated methods make it possible for both efficient and accurate protein structure modeling. The AI-aided modeling methods infer protein structures via machine learning and deep learning algorithms [11]. Currently, AI-driven methods like AlphaFold [12] are leading the research. In the near future, we can expect more accurate and efficient approaches for predicting protein structures.



Citation: Zhang, J.; Qian, J. Advances in Computational Intelligence-Based Methods of Structure and Function Prediction of Proteins. *Biomolecules* **2024**, *14*, 1083. <https://doi.org/10.3390/biom14091083>

Received: 22 August 2024

Accepted: 26 August 2024

Published: 29 August 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Table 1. Representative predictors of template-based, ab initio, and AI-aided modeling methods of protein structure. These research papers are sorted by the number of citations (scientific impact) in descending order.

Types of Methods	Representative Methods	Year	Citations	Reference
Template-Based Modeling	MODELLER	2016	10,180	[13]
	I-TASSER	2015	6060	[14]
	Sparks-X	2011	375	[15]
Ab Initio Modeling	QUARK	2012	1123	[16]
	RaptorX-Property	2016	574	[17]
	QMCPACK	2018	293	[18]
AI-Aided Modeling	AlphaFold	2021	25,578	[12]
	RaptorX_Contact	2019	417	[19]
	trRosetta	2021	404	[20]

In cells, all proteins have specific biological activities or functions, including structural supports and movements, enzymatic activities, and interactions with other ligands [21]. Protein functions depend on the corresponding 3D structures. For instance, Zhang et al. pointed out that the ligands tend to be located in the relatively small cavities of the protein surface [22]. If enzymes have buried active sites, substrates need to pass through the body of the protein in order to bind these sites [23]. According to recent studies [24–27], the approaches that predict protein function from structure include structure alignment, molecular docking, and AI-aided prediction. Researchers use Gene Ontology (GO) to detail the functions of a protein. The GO information includes biological process ontology, molecular function ontology, and cellular component ontology [28,29]. Figure 1 summarizes the protein relationships among sequence, structure, and function.

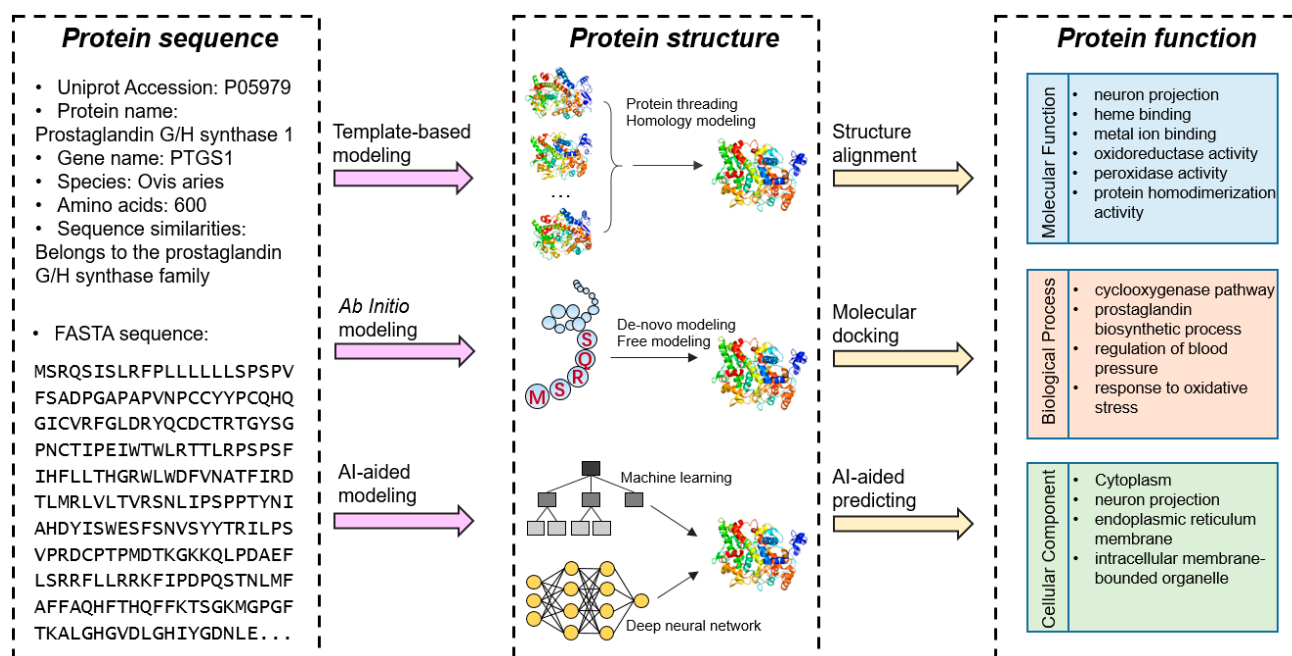


Figure 1. A schematic view of the protein sequence-structure-function relationships. Generally, protein sequence determines structure, and structure determines function.

This Special Issue consists of eight original research articles and one review of this topic. The research articles provide the readers with the latest developments in protein folding, secondary structure prediction, the HtrA protease family, SARS-CoV-2 spike variant complexes, heme distortion, Trp305, enzyme substrate promiscuity, and the hepatitis C virus genome. The review summarizes and compares existing network-based methods for predicting drug-disease associations.

Azulay et al. propose an interesting idea that compares and analyzes the similarity between proteins and origami [30]. Although the two things are not identical, they share some equivalences. Protein folding is driven by the physical arrangement of the residue chain and chemical forces. The corresponding crease patterns, like mountains or valleys, also appear in origami. The origami crease patterns and folding inspire scientists to explore protein folding constraints, properties of folded mediums, and folding energy. Besides that, the authors also discuss several unique mechanical properties, which have high stiffness-to-mass ratios and excellent abilities to withstand high forces. The capability of translating origami models to protein structures promises the visual design of de novo proteins and nanomaterials with the desired properties.

In [31], Guo et al. propose a novel method called CondGCNN to predict protein secondary structure. CondGCNN combines a conditionally parameterized convolutional network and a gated convolutional neural network. Particularly, the encoder layer of CondGCNN utilizes both the long short-term memory network and CondGCNN to compute protein sequential features. The results on benchmark testing datasets and a set of CASP datasets prove the good performance of their method.

Merski et al. investigate the structure-repeating module in the HtrA protease family [32]. They use a self-homology detection method based on a modified version of DOTTER [33] to analyze these protein sequences. As a result, they find a 26-residue segment/pattern, which forms an anti-parallel β -barrel structure. By using MUSCLE [34], the authors find that 13 out of 26 positions are evolutionary conserved. This repeating architecture has gone unnoticed, although these structures have been publicly available for two decades.

Verkhivker et al. perform a computational analysis of SARS-CoV-2 receptor-binding domain Omicron complexes with several ultra-potent antibodies [35]. They find that the dominant binding energy hotspots and allosteric centers of long-range interactions in the Omicron complexes share the same set of residues: Y449, Y453, L455, F486, Y489, and F490. These residues are conserved and hydrophobic with a low probability of mutation and are important for the receptor-binding domain and binding with the host receptor. If mutations occur in these residues, they will severely impair binding with the antibodies.

Compared to the isolated structure, heme in the host protein exhibits various degrees of distortion [36]. Moreover, the doming distortions in the oxygenated and deoxygenated states differ between hemoglobin and myoglobin [37]. In other words, the protein environment affects the heme molecular structure and controls the chemical properties of heme [22]. Kondo et al. construct a convolutional neural network to predict the distortion of heme from the 3D structure of the heme-binding pocket [38]. They examine the correlations between the shape of the cavity and the molecular structure of heme and obtain high correlation coefficients for saddling, ruffling, doming, and waving distortions.

The death-associated protein kinase (DAPK) family regulates important biological functions in human cells [39]. Among them, DAPK1 is the largest protein in its family and functions as a drug target in some diseases, such as cancer and Alzheimer's disease [40]. The research of Zhu et al. investigates the molecular mechanism of Trp305 (W305Y and W305D) in modulating DAPK1 activity [41]. They conclude that the W305D mutation enhanced the anti-correlated motions between DAPK1 and calcium/calmodulin, and the latter can interact with the W305Y DAPK1 mutant.

EP-Pred, a machine learning tool proposed by Xiang et al., is designed for identifying enzyme substrate promiscuity [42]. It adopts Possum [43] and iFeature [44] to extract evolutionary information and physicochemical properties, respectively. After feature

selection, the authors use SVM, KNN, and RidgeClassifier to construct an ensemble binary predictor. They use a hidden Markov approach to select promiscuous esterases from the Lipase Engineering Database. The EP-Pred confirms the validity of the selection and correctly recognizes all ten proteins.

Hepatitis C virus (HCV) infection is a major cause of liver failure and hepatocellular carcinoma worldwide [45]. In [46], Zhuang et al. aim to investigate how the stem-loop 1 and the 9th nucleotide of HCV affect the conformation and dynamics of the Ago2: miRNA: target RNA complex. They perform molecular dynamics simulations on the Ago2-miRNA complex and design a model wherein the Ago2 protein can adopt a more constrained conformation to protect the target RNA from dissociation. They find the mechanism of the Ago2-miRNA complex's protective effect on the HCV genome. This conclusion promises to offer guidance for the development of anti-HCV strategies.

The review by Kim et al. collects and compares the recently released network-based methods for drug-disease association prediction [47]. It categorizes these methods into three groups: graph mining, matrix factorization/matrix completion, and deep learning. The authors adopt two uniform datasets as the benchmark testing datasets. The two datasets are used to predict associations on the drug side and the disease side, respectively. Kim et al. compare the performance of the selected 11 predictors. Based on the experimental results, they find that the methods with graph mining and matrix factorization/matrix completion show better results than those with deep learning. Moreover, the current methods have higher accuracy on the drug side than on the disease side.

We hope that the readers will enjoy reading this Special Issue of Biomolecules and that the research articles on protein folding, secondary structure prediction, HtrA protease family, SARS-CoV-2 spike variant complexes, heme distortion, Trp305, enzyme substrate promiscuity, and hepatitis C virus genome, and the review on drug-disease associations will help advance the field and provide new ideas to researchers.

Author Contributions: Conceptualization, J.Z. and J.Q.; writing—original draft preparation, J.Q.; writing—review and editing, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Natural Science Foundation of Henan (Grant number 242300421410) and the Nanhu Scholars Program for Young Scholars of the Xinyang Normal University.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bereau, T.; Bachmann, M.; Deserno, M. Interplay between secondary and tertiary structure formation in protein folding cooperativity. *J. Am. Chem. Soc.* **2010**, *132*, 13129–13131. [[CrossRef](#)] [[PubMed](#)]
2. Kuhlman, B.; Bradley, P. Advances in protein structure prediction and design. *Nat. Rev. Mol. Cell Biol.* **2019**, *20*, 681–697. [[CrossRef](#)] [[PubMed](#)]
3. Hinz, U.; Consortium, U. From protein sequences to 3D-structures and beyond: The example of the UniProt knowledgebase. *Cell. Mol. Life Sci.* **2010**, *67*, 1049–1064. [[CrossRef](#)]
4. Schneider, T.; Riedel, K. Environmental proteomics: Analysis of structure and function of microbial communities. *Proteomics* **2010**, *10*, 785–798. [[CrossRef](#)]
5. Zhang, J.; Basu, S.; Kurgan, L. HybridDBRpred: Improved sequence-based prediction of DNA-binding amino acids using annotations from structured complexes and disordered proteins. *Nucleic Acids Res.* **2024**, *52*, e10. [[CrossRef](#)]
6. Kryshtafovych, A.; Schwede, T.; Topf, M.; Fidelis, K.; Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round XV. *Proteins Struct. Funct. Bioinform.* **2023**, *91*, 1539–1549. [[CrossRef](#)]
7. Simpkin, A.J.; Mesdaghi, S.; Sánchez Rodríguez, F.; Elliott, L.; Murphy, D.L.; Kryshtafovych, A.; Keegan, R.M.; Rigden, D.J. Tertiary structure assessment at CASP15. *Proteins Struct. Funct. Bioinform.* **2023**, *91*, 1616–1635. [[CrossRef](#)] [[PubMed](#)]
8. Ye, N.; Zhou, F.; Liang, X.; Chai, H.; Fan, J.; Li, B.; Zhang, J. A Comprehensive Review of Computation-Based Metal-Binding Prediction Approaches at the Residue Level. *BioMed Res. Int.* **2022**, *2022*, 8965712. [[CrossRef](#)] [[PubMed](#)]
9. Fiser, A. Template-based protein structure modeling. *Comput. Biol.* **2010**, *673*, 73–94.
10. Wu, F.; Xu, J. Deep template-based protein structure prediction. *PLoS Comput. Biol.* **2021**, *17*, e1008954. [[CrossRef](#)]
11. Kumar, H.; Kim, P. Artificial intelligence in fusion protein three-dimensional structure prediction: Review and perspective. *Clin. Transl. Med.* **2024**, *14*, e1789. [[CrossRef](#)] [[PubMed](#)]

12. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Židek, A.; Potapenko, A. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Webb, B.; Sali, A. Comparative protein structure modeling using MODELLER. *Curr. Protoc. Bioinform.* **2016**, *54*, 5.6.1–5.6.37. [\[CrossRef\]](#)
14. Yang, J.; Yan, R.; Roy, A.; Xu, D.; Poisson, J.; Zhang, Y. The I-TASSER Suite: Protein structure and function prediction. *Nat. Methods* **2015**, *12*, 7–8. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Yang, Y.; Faraggi, E.; Zhao, H.; Zhou, Y. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics* **2011**, *27*, 2076–2082. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Xu, D.; Zhang, Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins Struct. Funct. Bioinform.* **2012**, *80*, 1715–1735. [\[CrossRef\]](#)
17. Wang, S.; Li, W.; Liu, S.; Xu, J. RaptorX-Property: A web server for protein structure property prediction. *Nucleic Acids Res.* **2016**, *44*, W430–W435. [\[CrossRef\]](#)
18. Kim, J.; Baczewski, A.D.; Beaudet, T.D.; Benali, A.; Bennett, M.C.; Berrill, M.A.; Blunt, N.S.; Borda, E.J.L.; Casula, M.; Ceperley, D.M. QMCPACK: An open source ab initio quantum Monte Carlo package for the electronic structure of atoms, molecules and solids. *J. Phys. Condens. Matter* **2018**, *30*, 195901. [\[CrossRef\]](#)
19. Xu, J. Distance-based protein folding powered by deep learning. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 16856–16865. [\[CrossRef\]](#)
20. Du, Z.; Su, H.; Wang, W.; Ye, L.; Wei, H.; Peng, Z.; Anishchenko, I.; Baker, D.; Yang, J. The trRosetta server for fast and accurate protein structure prediction. *Nat. Protoc.* **2021**, *16*, 5634–5651. [\[CrossRef\]](#)
21. Zhang, J.; Ghadermarzi, S.; Kurgan, L. Prediction of protein-binding residues: Dichotomy of sequence-based methods developed using structured complexes versus disordered proteins. *Bioinformatics* **2020**, *36*, 4729–4738. [\[CrossRef\]](#)
22. Zhang, J.; Chai, H.; Gao, B.; Yang, G.; Ma, Z. HEMESPred: Structure-Based Ligand-Specific Heme Binding Residues Prediction by Using Fast-Adaptive Ensemble Learning Scheme. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2018**, *15*, 147–156. [\[CrossRef\]](#)
23. Kingsley, L.J.; Lill, M.A. Substrate tunnels in enzymes: Structure-function relationships and computational methodology. *Proteins* **2015**, *83*, 599–611. [\[CrossRef\]](#)
24. Huang, J.; Lin, Q.; Fei, H.; He, Z.; Xu, H.; Li, Y.; Qu, K.; Han, P.; Gao, Q.; Li, B.; et al. Discovery of deaminase functions by structure-based protein clustering. *Cell* **2023**, *186*, 3182–3195.e3114. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Yang, C.; Chen, E.A.; Zhang, Y. Protein-Ligand Docking in the Machine-Learning Era. *Molecules* **2022**, *27*, 4568. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Walder, M.; Edelstein, E.; Carroll, M.; Lazarev, S.; Fajardo, J.E.; Fiser, A.; Viswanathan, R. Integrated structure-based protein interface prediction. *BMC Bioinform.* **2022**, *23*, 301. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Gligorićević, V.; Renfrew, P.D.; Kosciółek, T.; Leman, J.K.; Berenberg, D.; Vatanen, T.; Chandler, C.; Taylor, B.C.; Fisk, I.M.; Vlamakis, H.; et al. Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.* **2021**, *12*, 3168. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Blake, J.A.; Dolan, M.; Drabkin, H.; Hill, D.P.; Li, N.; Sitnikov, D.; Bridges, S.; Burgess, S.; Buza, T.; McCarthy, F.; et al. Gene Ontology annotations and resources. *Nucleic Acids Res.* **2013**, *41*, D530–D535.
29. Zhang, J.; Ghadermarzi, S.; Katuwawala, A.; Kurgan, L. DNAGenie: Accurate prediction of DNA-type-specific binding residues in protein sequences. *Brief. Bioinform.* **2021**, *22*, bbab336. [\[CrossRef\]](#)
30. Azulay, H.; Lutaty, A.; Qvit, N. How Similar Are Proteins and Origami? *Biomolecules* **2022**, *12*, 622. [\[CrossRef\]](#)
31. Guo, Y.; Wu, J.; Ma, H.; Wang, S.; Huang, J. Deep ensemble learning with atrous spatial pyramid networks for protein secondary structure prediction. *Biomolecules* **2022**, *12*, 774. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Merski, M.; Macedo-Ribeiro, S.; Wiczorek, R.M.; Górna, M.W. The Repeating, Modular Architecture of the HtrA Proteases. *Biomolecules* **2022**, *12*, 793. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Merski, M.; Młynarczyk, K.; Ludwiczak, J.; Skrzeczkowski, J.; Dunin-Horkawicz, S.; Górna, M.W. Self-analysis of repeat proteins reveals evolutionarily conserved patterns. *BMC Bioinform.* **2020**, *21*, 179. [\[CrossRef\]](#)
34. Madeira, F.; Park, Y.M.; Lee, J.; Buso, N.; Gur, T.; Madhusoodanan, N.; Basutkar, P.; Tivey, A.R.N.; Potter, S.C.; Finn, R.D.; et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* **2019**, *47*, W636–W641. [\[CrossRef\]](#)
35. Verkhivker, G.; Agajanian, S.; Kassab, R.; Krishnan, K. Integrating Conformational Dynamics and Perturbation-Based Network Modeling for Mutational Profiling of Binding and Allostery in the SARS-CoV-2 Spike Variant Complexes with Antibodies: Balancing Local and Global Determinants of Mutational Escape Mechanisms. *Biomolecules* **2022**, *12*, 964. [\[CrossRef\]](#)
36. Kondo, H.X.; Kanematsu, Y.; Masumoto, G.; Takano, Y. PyDISH: Database and analysis tools for heme porphyrin distortion in heme proteins. *Database* **2023**, *2023*, baaa066. [\[CrossRef\]](#) [\[PubMed\]](#)
37. Kondo, H.X.; Takano, Y. Analysis of Fluctuation in the Heme-Binding Pocket and Heme Distortion in Hemoglobin and Myoglobin. *Life* **2022**, *12*, 210. [\[CrossRef\]](#)
38. Kondo, H.X.; Iizuka, H.; Masumoto, G.; Kabaya, Y.; Kanematsu, Y.; Takano, Y. Elucidation of the Correlation between Heme Distortion and Tertiary Structure of the Heme-Binding Pocket Using a Convolutional Neural Network. *Biomolecules* **2022**, *12*, 1172. [\[CrossRef\]](#)
39. Farag, A.K.; Roh, E.J. Death-associated protein kinase (DAPK) family modulators: Current and future therapeutic outcomes. *Med. Res. Rev.* **2019**, *39*, 349–385. [\[CrossRef\]](#)

40. Chen, D.; Zhou, X.Z.; Lee, T.H. Death-Associated Protein Kinase 1 as a Promising Drug Target in Cancer and Alzheimer's Disease. *Recent Pat. Anti-Cancer Drug Discov.* **2019**, *14*, 144–157. [[CrossRef](#)]
41. Zhu, Y.P.; Gao, X.Y.; Xu, G.H.; Qin, Z.F.; Ju, H.X.; Li, D.C.; Ma, D.N. Computational Dissection of the Role of Trp305 in the Regulation of the Death-Associated Protein Kinase-Calmodulin Interaction. *Biomolecules* **2022**, *12*, 1395. [[CrossRef](#)] [[PubMed](#)]
42. Xiang, R.; Fernandez-Lopez, L.; Robles-Martín, A.; Ferrer, M.; Guallar, V. EP-Pred: A Machine Learning Tool for Bioprospecting Promiscuous Ester Hydrolases. *Biomolecules* **2022**, *12*, 1529. [[CrossRef](#)] [[PubMed](#)]
43. Wang, J.; Yang, B.; Revote, J.; Leier, A.; Marquez-Lago, T.T.; Webb, G.; Song, J.; Chou, K.C.; Lithgow, T. POSSUM: A bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles. *Bioinformatics* **2017**, *33*, 2756–2758. [[CrossRef](#)]
44. Chen, Z.; Zhao, P.; Li, F.; Leier, A.; Marquez-Lago, T.T.; Wang, Y.; Webb, G.I.; Smith, A.I.; Daly, R.J.; Chou, K.C.; et al. iFeature: A Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* **2018**, *34*, 2499–2502. [[CrossRef](#)]
45. Gower, E.; Estes, C.; Blach, S.; Razavi-Shearer, K.; Razavi, H. Global epidemiology and genotype distribution of the hepatitis C virus infection. *J. Hepatol.* **2014**, *61*, S45–S57. [[CrossRef](#)] [[PubMed](#)]
46. Zhuang, H.; Ji, D.; Fan, J.; Li, M.; Tao, R.; Du, K.; Lu, S.; Chai, Z.; Fan, X. Mechanistic Insights into the Protection Effect of Argonaute-RNA Complex on the HCV Genome. *Biomolecules* **2022**, *12*, 1631. [[CrossRef](#)]
47. Kim, Y.; Jung, Y.S.; Park, J.H.; Kim, S.J.; Cho, Y.R. Drug-Disease Association Prediction Using Heterogeneous Networks for Computational Drug Repositioning. *Biomolecules* **2022**, *12*, 1497. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.