DEEP LEARNING IN MODELLING THE PROTEIN–LIGAND INTERACTION: NEW PATHWAYS IN DRUG DEVELOPMENT

Barykin AD[1,2], Chepurnykh TV[1], Osipova ZM[1,3] ✉

[1] Shemyakin–Ovchinnikov Institute of Bioorganic Chemistry of the Russian Academy of Sciences, Moscow, Russia

[2] Moscow Institute of Physics and Technology (MIPT), Dolgoprudny, Russia

[3] Pirogov Russian National Research Medical University, Moscow, Russia

The deep learning technologies have become the driver of the revolutionary changes in scientific research in various fields. The AlphaFold-2 neural network software development that has solved the semicentennial problem of 3D protein structure prediction based on primary amino acid sequence is the most obvious example of using such technologies in structural biology and biomedicine. The use of deep learning methods for the prediction of protein–ligand interactions can considerably simplify predicting, speed up the development of new effective pharmaceuticals and change the concept of drug design.

Keywords: docking, protein–ligand interaction, neural networks, deep learning

Funding: the study was supported by the Russian Science Foundation grant, project № 22-44-02024 (https://rscf.ru/project/22-44-02024/).

Author contribution: Barykin AD — literature review, manuscript writing, Chepurnykh TV — concept, literature review, manuscript writing and editing, Osipova ZM — project management, manuscript editing.

✉ **Correspondence should be addressed:** Zinaida M. Osipova
Miklukho-Maklaya, 16/10, Moscow, 117997, Russia; zkaskova@ibch.ru

Received: 06.12.2023 **Accepted:** 22.01.2024 **Published online:** 08.02.2024

DOI: 10.24075/brsmu.2024.002

ГЛУБОКОЕ ОБУЧЕНИЕ В МОДЕЛИРОВАНИИ БЕЛОК-ЛИГАНДНОГО ВЗАИМОДЕЙСТВИЯ: НОВЫЕ ПУТИ В РАЗРАБОТКЕ ЛЕКАРСТВЕННЫХ ПРЕПАРАТОВ

А. Д. Барыкин[1,2], Т. В. Чепурных[1], З. М. Осипова[1,3] ✉

[1] Институт биоорганической химии имени М. М. Шемякина и Ю. А. Овчинникова Российской академии наук, Москва, Россия

[2] Московский физико-технический институт, Долгопрудный, Россия

[3] Российский национальный исследовательский медицинский университет имени Н. И. Пирогова, Москва, Россия

Технологии глубокого обучения стали драйвером революционных изменений в научных исследованиях разных областей. Наиболее ярким примером их применения в области структурной биологии и биомедицины является программная разработка нейросеть AlphaFold-2, решившая полувековую проблему предсказания 3D-структуры белков по первичной аминокислотной последовательности. Использование методов глубокого обучения для предсказания белок-лигандных взаимодействий сможет значительно упростить предсказание, ускорить разработку новых эффективных лекарственных препаратов и поменять концепцию драг-дизайна.

Ключевые слова: докинг, белок-лигандное взаимодействие, алгоритм нейросети, глубокое обучение

Финансирование: исследование выполнено за счет гранта Российского научного фонда № 22-44-02024, https://rscf.ru/project/22-44-02024/.

Вклад авторов: А. Д. Барыкин — анализ литературы, написание рукописи, Т. В. Чепурных — идея, анализ литературы, написание и редактирование рукописи, З. М. Осипова — руководство проектом, редактирование рукописи.

✉ **Для корреспонденции:** Зинаида Михайловна Осипова
ул. Миклухо-Маклая, 16/10, г. Москва, 117997, Россия; zkaskova@ibch.ru

Статья получена: 06.12.2023 **Статья принята к печати:** 22.01.2024 **Опубликована онлайн:** 08.02.2024

DOI: 10.24075/vrgmu.2024.002

Computer (*in silico*) modelling of protein–ligand interaction plays a key role in biomedical research and represents one of the fundamental challenges of the ongoing process of developing new pharmaceuticals. The higher is the affinity and selectivity of the bioactive molecule binding with the receptor or enzyme, the more effective and safe the resulting drug candidate will be. The model reliability determines the quantity and quality of the candidate molecules, which will undergo the expensive procedure of chemical synthesis and testing in vitro and in vivo. Modelling often represents the key phase: this to the great extent determines the time and cost of the design, as well as the final price of the drug [1]. Until recently, there was no highly effective method for automated bioinformatics analysis of the protein–ligand interaction.

Conventional computer modelling methods

Molecular docking is a molecular modelling method predicting the best position of the ligand relative to the target protein that uses their 3D structures and the molecular interaction energy scoring functions. The scoring function learning (Fig. 1A) is usually based on the set of the experimentally determined affinity values of protein binding to the ligands similar to the studied one. Thus, the prediction accuracy will depend directly on the similarity of the studied new candidate to the known ligands from the database.

The great diversity of scoring functions can be explained by insufficient reliability of each of these functions in case of solving a specific problem. Various scoring functions are more appropriate for various classes of ligands, however, there is no absolute guarantee of the result even when the method is selected correctly. That is why consensus assessment (using the data of several scoring functions at once) increases the likelihood of successful docking [2].

In case of rigid docking, the algorithms consider the ligand and target molecules as solids, while in case of dynamic docking the programs admit the possibility of conformational changes in the ligand associated with binding. The methods

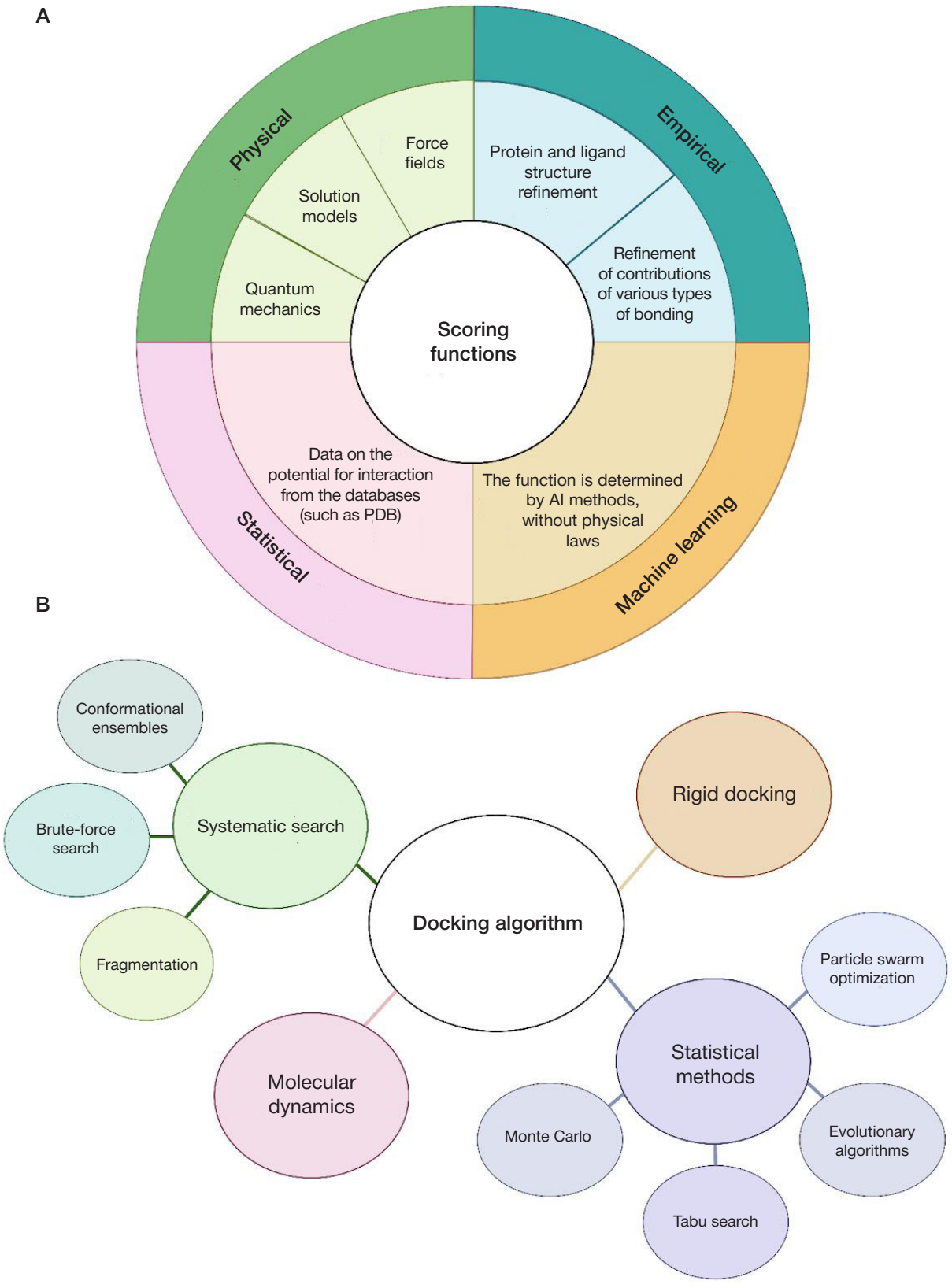

Fig. 1. Molecular docking algorithms. **A**. Types of scoring functions for molecular docking. **B**. Variants of molecular docking algorithms and molecular dynamics

underlying the docking algorithms (Fig. 1B) can be divided roughly into systematic and statistical. The systematic methods split molecule into several parts, thereby allowing one to assess affinity of interaction for each part, and then ensure covalent cross-linking of the parts in order to reassemble the ligand. Statistical methods to find the global energy minimum generate random changes, for each of which the thermodynamic state is estimated [3]. Statistical methods include the Monte Carlo method, tabu search, particle swarm optimization, and evolutionary algorithms. The systematic algorithms guarantee achievement of results in a finite number of steps (usually very large), while the statistical ones can “miss” the energy state of interest. However, in practice, statistical algorithms more often yield more reliable results than systematic ones.

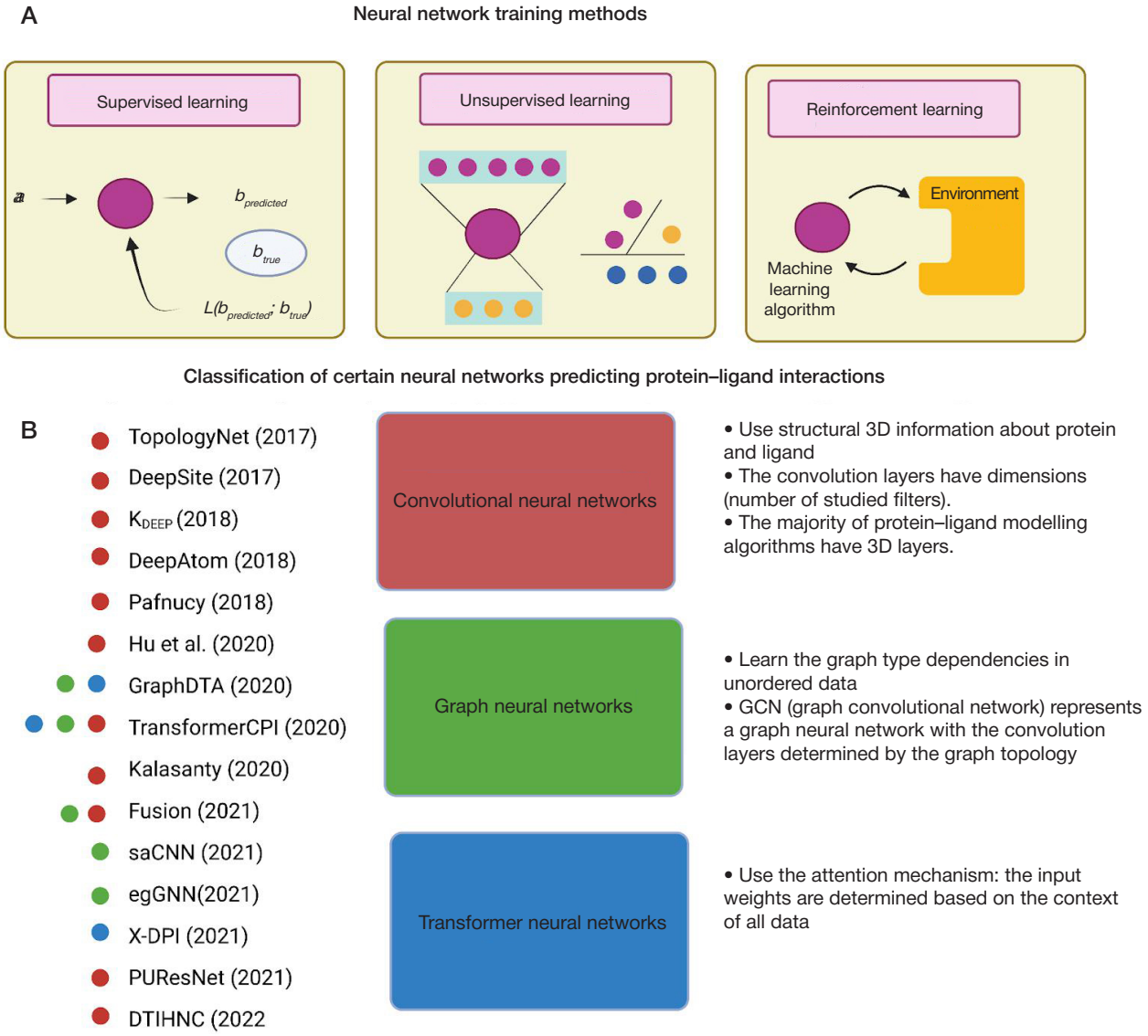

Fig. 2. Neural network operation algorithms. **A**. Neural network training methods. **B**. Major types of neural networks predicting the protein–ligand interactions

In the last two decades, dozens of free and commercially available molecular docking programs have emerged: DOCK, AutoDock, Surflex, LigandFit, MCDock, LeDock, AutoDock Vina, rDock, UCSF Dock, and many more [4]. The programs usually use several algorithms at once, allowing one to adapt docking for specific enzyme–ligand pairs.

In the majority of cases, the modern protein–ligand docking methods correctly determine the ligand binding site and binding mechanism, however, these are unable to define its affinity accurately enough [5]. This significantly reduces the method applicability for discovery of new pharmaceuticals, since the candidate molecules are selected based specifically on the binding energy.

The molecular dynamics (MD) method is based on the use of the equations of atom motion and the empirical potential energy functions for calculation of interatomic interactions and characterization of the molecular system evolution over time. The interatomic interactions include elastic interactions (corresponding to covalent bonds) and Van der Waals forces. The most important post-processing methods to calculate free energy of the bond in the protein–ligand interaction complex also use the principles of molecular mechanics involving the Poisson–Boltzmann equation / Generalized Born model, as well as some additional approaches, such as thermodynamic integration and free analysis [6].

The molecular trajectory length determined by the number of simulation steps is the main limitation of the molecular dynamics method. The simulation time step should be comparable with the fastest motion in the system, i.e. with the bond fluctuations (1–2 fs). Thus, modelling slow processes, such as large domain motions and binding (μs–ms), requires many MD steps, which significantly increases the amount of computation. That is why detection of actual protein–ligand binding is a very rare phenomenon [7]. It was expected that the MD modelling based on the binding computation involving the use of molecular mechanics and the Poisson–Boltzmann equation would contribute significantly to the solution of real problems, such as identification of the most advantageous combination for the protein–ligand pairs with their further optimization.

Deep learning: another chapter in modelling the protein–ligand interaction

The first deep learning methods emerged in the mid-1960s, however, these became popular by the mid-2000s due to the increase in processing power and the emergence of large experimental data sets. Today, application of deep learning technologies to problems in various disciplines has yielded the

results that are not inferior, and sometimes even superior, to the results yielded by conventional methods. Creation of the AlphaFold-2 algorithm predicting the protein tertiary structure based on the primary amino acid sequence within only a few minutes is the most obvious example [8], which has become a revolution in the field of structural biology.

The deep learning models were proposed for the prediction of protein–ligand interaction as an alternative to conventional docking based on the search for the free energy minimum [9]. The advantage of deep learning is that there is a possibility of studying the protein–ligand interaction based directly on the spatial arrangement of atoms, without selecting the mathematical parameters that not always reflect the actual binding mechanism. This method to predict protein–ligand interaction is currently developing rapidly: the DeepSite neural network model published in 2017 [10] correctly determined 23.8% of ligand binding sites in the specific dataset, while the Kalasanty neural network published in 2020 yielded the result of 44.6% for the same sample. In 2021, PUResNet significantly improved the prediction results (53% success for PUResNet vs. 51% for Kalasanty) [11].

To date, a wide variety of the neural network operation and training algorithms has been developed (Fig. 2). Convolutional neural networks (CNN), graph neural networks (GNN), and transformer neural networks are usually used to study protein–ligand interactions. Convolutional neural networks consider the relationships between pairs of atoms through the prism of their relative spatial positions. The operation principle of graph neural networks is based on recording the threshold values aimed to determine the type of interaction between atoms (covalent or non-covalent). The use of the lower number of parameters represents the potential advantage of this approach. The combinations of several algorithms are also used, or the other modules (such as denoising autoencoder) that improve the outcome are added [12].

CONCLUSION

No definite leader has been revealed among various neural network architectures: accuracy of the results yielded by each algorithm depends on the types of proteins and ligands, affinity, and binding mechanism. According to the latest data, the egGNN and saCNN graph neural networks turned out to be the most successful in terms of ligand affinity prediction [13, 14], however, the differences from convolutional neural networks are not critical. We believe that this is due to the fact that no optimal prediction algorithm has yet been developed. Creating such an algorithm, judging by the pace of AI development in computational biology, is more likely to take years, not decades. Just like AlphaFold-2 changed the paradigm in the area of protein structure investigation in 2020, the use of artificial intelligence in biomedical research opens a new chapter in pharmaceutical industry and drug design.

The prospects of using AI for drug discovery have become apparent for the industry, since the use of AI significantly accelerates and reduces the cost of the conventional 12-year lifecycle of drug development. Over the past five years, almost all large pharmaceutical companies announced a partnership with the leading AI companies (Sanofi — Aily Labs, Pfizer — IBM, Novartis — Microsoft, AstraZeneca — Benevolent, etc.). The research details are likely to be protected by trade secret for a long time, however, press releases are being regularly issued that suggest introduction of deep learning into the ongoing R&D processes. Furthermore, more and more reports appear of the success of drug candidates designed using AI that are being prepared for or are through clinical trials. The examples include halicin (the promising broad-spectrum antibiotic, preclinical trial) [15], INS018_055 (drug for idiopathic pulmonary fibrosis, phase 2 clinical trial), REC-2282, REC-994, REC-4881, BEN-2293, EXS-21546, RLY-4008, EXS-4318, BEN-8744, etc. [16].

We believe that the search for the pool of new bioactive molecules will accelerate dramatically in the near future, and deep learning will become an essential element of the process of developing new pharmaceuticals. However, competent selection of the data that are used to train a neural network model still represents one of the topical problems on the path to the widespread use of deep learning for drug discovery, since the quality of these data is critical for reliability of predictions made by the model. In this regard, the task of the most effective training involving the use of incomplete or small datasets remains the main challenge faced by AI in the field of drug design [17].

References

1. Lazo JS. Rear-view mirrors and crystal balls: a brief reflection on drug discovery. Mol Interv. 2008; 8 (2): 60–3. DOI: 10.1124/mi.8.2.1.
2. Blanes-Mira C, Fernández-Aguado P, de Andrés-López J, Fernández-Carvajal A, Ferrer-Montiel A, Fernández-Ballester G. Comprehensive survey of consensus docking for high-throughput virtual screening. Molecules. 2023; 28 (1): 175. DOI: 10.3390/molecules28010175.
3. Zhao L, Zhu Y, Wang J, Wen N, Wang C, Cheng L. A brief review of protein–ligand interaction prediction. Comput Struct Biotechnol J. 2022; 20: 2831–8. DOI: 10.1016/j.csbj.2022.06.004.
4. Pagadala NS, Syed K, Tuszynski J. Software for molecular docking: a review. Biophys Rev. 2017; 9: 91–102. DOI: 10.1007/s12551-016-0247-1. PubMed PMID: 28510083.
5. Wang Z, Sun H, Yao X, Li D, Xu L, Li Y, et al. Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power. Phys Chem Chem Phys. 2016; 18: 12964–75. DOI: 10.1039/c6cp01555g.
6. Perez JJ, Perez RA, Perez A. Computational modeling as a tool to investigate PPI: from drug design to tissue engineering. Front Mol Biosci. 2021; 8: 681617. DOI: 10.3389/fmolb.2021.681617. PubMed PMID: 34095231.
7. Santos LHS, Ferreira RS, Caffarena ER. Integrating molecular docking and molecular dynamics simulations. Methods Mol Biol. 2019; 2053: 13–34. DOI: 10.3389/fmolb.2021.681617.
8. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021; 596: 583–9. DOI: 10.1038/s41586-021-03819-2.
9. Crampon K, Giorkallos A, Deldossi M, Baud S, Steffenel LA. Machine-learning methods for ligand–protein molecular docking. Drug Discov Today. 2021; 27(1): 151–64. DOI: 10.1016/j.drudis.2021.09.007.
10. Jiménez J, Doerr S, Martínez-Rosell G, Rose AS, De Fabritiis G. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. Bioinformatics. 2017; 33 (19): 3036–42. DOI: 10.1093/bioinformatics/btx350.
11. Kandel J, Tayara H, Chong KT. PUResNet: prediction of protein-ligand binding sites using deep residual neural network. J Cheminform. 2021; 13: 65. DOI: 10.1186/s13321-021-00547-7.
12. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. Nat Rev Mol Cell Biol. 2022; 23: 40–55. DOI: 10.1038/s41580-021-00407-0.

13. Wang Y, Jiao Q, Wang J, Cai X, Zhao W, Cui X. Prediction of protein-ligand binding affinity with deep learning. Comput Struct Biotechnol J. 2023; 21: 5796–806. DOI: 10.1016/j.csbj.2023.11.009.
14. Zhao L, Zhu Y, Wang J, Wen N, Wang C, Cheng L. A brief review of protein–ligand interaction prediction. Comput Struct Biotechnol J. 2022; 20: 2831–8. DOI: 10.1016/j.csbj.2022.06.004.
15. Stokes JM, Yang K,Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, et al. A deep learning approach to antibiotic discovery. Cell. 2020; 180: P688–702.E13. DOI: 10.1016/j.cell.2020.01.021.
16. Arnold C. Inside the nascent industry of AI-designed drugs. Nat Med. 2023; 29: 1292–5. DOI: 10.1038/s41591-023-02361-0.
17. Yang B, Li K, Zhong X, Zou J. Implementation of deep learning in drug design. MedComm — Fut Med. 2022; 1: e18. DOI: 10.1002/mef2.18.

Литература

1. Lazo JS. Rear-view mirrors and crystal balls: a brief reflection on drug discovery. Mol Interv. 2008; 8 (2): 60–3. DOI: 10.1124/mi.8.2.1.
2. Blanes-Mira C, Fernández-Aguado P, de Andrés-López J, Fernández-Carvajal A, Ferrer-Montiel A, Fernández-Ballester G. Comprehensive survey of consensus docking for high-throughput virtual screening. Molecules. 2023; 28 (1): 175. DOI: 10.3390/molecules28010175.
3. Zhao L, Zhu Y, Wang J, Wen N, Wang C, Cheng L. A brief review of protein–ligand interaction prediction. Comput Struct Biotechnol J. 2022; 20: 2831–8. DOI: 10.1016/j.csbj.2022.06.004.
4. Pagadala NS, Syed K, Tuszynski J. Software for molecular docking: a review. Biophys Rev. 2017; 9: 91–102. DOI: 10.1007/s12551-016-0247-1. PubMed PMID: 28510083.
5. Wang Z, Sun H, Yao X, Li D, Xu L, Li Y, et al. Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power. Phys Chem Chem Phys. 2016; 18: 12964–75. DOI: 10.1039/c6cp01555g.
6. Perez JJ, Perez RA, Perez A. Computational modeling as a tool to investigate PPI: from drug design to tissue engineering. Front Mol Biosci. 2021; 8: 681617. DOI: 10.3389/fmolb.2021.681617. PubMed PMID: 34095231.
7. Santos LHS, Ferreira RS, Caffarena ER. Integrating molecular docking and molecular dynamics simulations. Methods Mol Biol. 2019; 2053: 13–34. DOI: 10.3389/fmolb.2021.681617.
8. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021; 596: 583–9. DOI: 10.1038/s41586-021-03819-2.
9. Crampon K, Giorkallos A, Deldossi M, Baud S, Steffenel LA. Machine-learning methods for ligand–protein molecular docking. Drug Discov Today. 2021; 27(1): 151–64. DOI: 10.1016/j.drudis.2021.09.007.
10. Jiménez J, Doerr S, Martínez-Rosell G, Rose AS, De Fabritiis G. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. Bioinformatics. 2017; 33 (19): 3036–42. DOI: 10.1093/bioinformatics/btx350.
11. Kandel J, Tayara H, Chong KT. PUResNet: prediction of protein-ligand binding sites using deep residual neural network. J Cheminform. 2021; 13: 65. DOI: 10.1186/s13321-021-00547-7.
12. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. Nat Rev Mol Cell Biol. 2022; 23: 40–55. DOI: 10.1038/s41580-021-00407-0.
13. Wang Y, Jiao Q, Wang J, Cai X, Zhao W, Cui X. Prediction of protein-ligand binding affinity with deep learning. Comput Struct Biotechnol J. 2023; 21: 5796–806. DOI: 10.1016/j.csbj.2023.11.009.
14. Zhao L, Zhu Y, Wang J, Wen N, Wang C, Cheng L. A brief review of protein–ligand interaction prediction. Comput Struct Biotechnol J. 2022; 20: 2831–8. DOI: 10.1016/j.csbj.2022.06.004.
15. Stokes JM, Yang K,Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, et al. A deep learning approach to antibiotic discovery. Cell. 2020; 180: P688–702.E13. DOI: 10.1016/j.cell.2020.01.021.
16. Arnold C. Inside the nascent industry of AI-designed drugs. Nat Med. 2023; 29: 1292–5. DOI: 10.1038/s41591-023-02361-0.
17. Yang B, Li K, Zhong X, Zou J. Implementation of deep learning in drug design. MedComm — Fut Med. 2022; 1: e18. DOI: 10.1002/mef2.18.