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

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

Компьютерное (*in silico*) моделирование белок-лигандного взаимодействия играет ключевую роль в биомедицинских исследованиях и является одной из фундаментальных задач в современном процессе разработки новых лекарственных препаратов. Чем более аффинно и избирательно биоактивная молекула связывается с рецептором или ферментом, тем более эффективным и безопасным будет итоговый лекарственный кандидат. Достоверность моделирования определяет количество и качество молекул-кандидатов, которые будут проходить дорогостоящую процедуру химического синтеза и испытаний *in vitro* и *in vivo*. Стадия моделирования часто является ключевой: от нее в большой степени будут зависеть время, стоимость разработки и конечная цена лекарства [1]. Высокоэффективного метода биоинформатической автоматизированной оценки белок-лигандного взаимодействия до недавнего времени не существовало.

Классические методы компьютерного моделирования

Молекулярный докинг — метод молекулярного моделирования, предсказывающий наилучшее положение лиганда относительно белка-мишени, который использует их трехмерные структуры и оценочные функции энергии взаимодействия молекул (scoring functions). Обучение оценочных функций (рис. 1А) обычно происходит на основе набора экспериментально определенной аффинности связывания белка с лигандами, похожими на изучаемый. Правильность прогноза, таким образом, будет напрямую зависеть от степени сходства нового изучаемого кандидата и известных лигандов из базы данных.

Большое разнообразие оценочных функций можно объяснить недостаточной надежностью каждой из них в случае решения конкретной задачи. Разные оценочные функции лучше подходят для разных классов лигандов, но даже в случае правильного подбора метода не существует

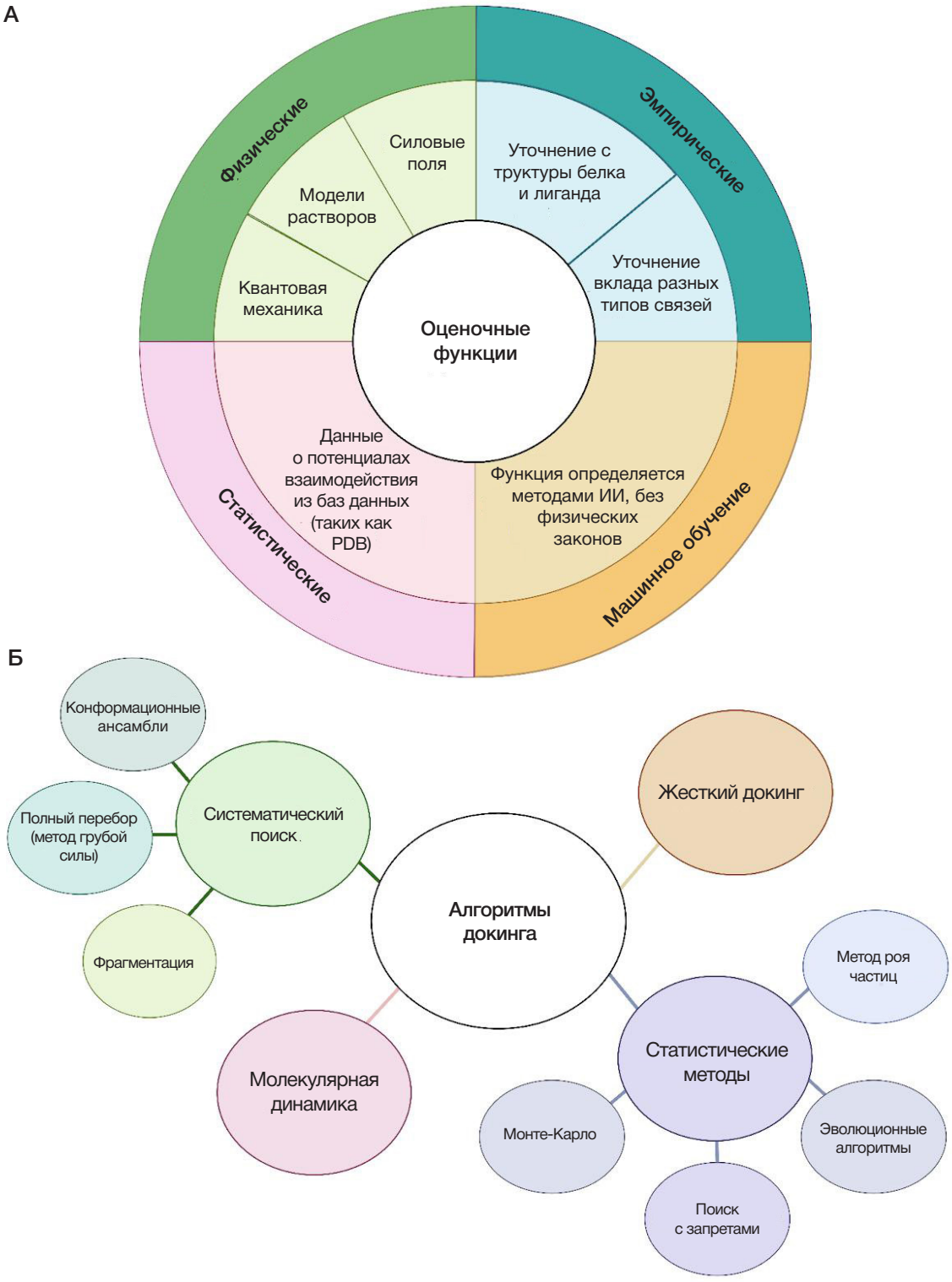

Рис. 1. Алгоритмы молекулярного докинга. **А**. Виды оценочных функций для молекулярного докинга. **Б**. Варианты алгоритмов молекулярного докинга и молекулярная динамика

абсолютной гарантии результата. Поэтому консенсусное оценивание (использование данных сразу нескольких оценочных функций) повышает вероятность успеха докинга [2].

В случае «жесткого» докинга алгоритмы рассматривают молекулы лиганда и мишени как твердые тела, в случае «динамического» докинга программы допускают возможность конформационных изменений в лиганде при его связывании. Методы, лежащие в основе алгоритмов докинга (рис. 1Б), можно условно разделить на систематические и статистические. Систематические методы разбивают молекулу лиганда на несколько частей, что позволяет оценивать аффинность взаимодействия каждой части, а затем части ковалентно «сшивают», чтобы «пересобрать» лиганд. Статистические методы для поиска глобального минимума энергии генерируют случайные изменения, для каждого из которых оценивается термодинамическое состояние [3]. К статистическим относят метод Монте-Карло, поиск с запретами, метод «роя частиц» и эволюционные алгоритмы. Систематические

алгоритмы гарантируют достижение результата за конечное число шагов (как правило, очень большое), статистические могут «пропустить» искомое энергетическое состояние. Однако на практике статистические алгоритмы часто показывают более достоверные результаты, чем систематические.

За последние два десятилетия возникли десятки бесплатных и коммерческих программ для молекулярного докинга: DOCK, AutoDock, Surflex, LigandFit, MCDock, LeDock, AutoDock Vina, rDock, UCSF Dock и многие другие [4]. Обычно программы используют сразу несколько алгоритмов, позволяя адаптировать докинг под конкретные пары фермент–лиганд.

В большинстве случаев методы современного белок-лигандного докинга верно определяют сайт и механизм связывания лиганда, но не могут установить его аффинность с достаточной точностью [5]. Это значительно снижает применимость метода для поиска новых лекарственных средств, поскольку подбор молекул-кандидатов осуществляется именно по величине энергии связывания.

Метод молекулярной динамики (МД) основан на использовании уравнений движения атомов и эмпирических функций потенциальной энергии для расчета межатомных взаимодействий и описания эволюции молекулярной системы во времени. Взаимодействия между атомами включают упругие взаимодействия (соответствующие ковалентным химическим связям) и силы Ван-дер-Ваальса. Наиболее важные методы постобработки для расчета свободной энергии связи комплекса взаимодействия белка и лиганда используют также принципы молекулярной механики с применением уравнения Пуассона–Больцмана / обобщенной модели Борна, а также дополнительные подходы, такие как термодинамическая интеграция и свободный анализ [6].

Основным ограничением в методе молекулярной динамики является длина молекулярной траектории, которая задается количеством шагов симуляции. Временной шаг симуляции должен быть сопоставим с самым быстрыми движениями в системе, а именно колебаниями связей (1–2 фс). Таким образом, для моделирования медленных процессов, например движения больших доменов и связывания (мкс–мс), требуется большое число шагов МД, что значительно увеличивает объем вычислений. Поэтому наблюдение фактического связывания белка с лигандом — очень редкое явление [7]. Ожидалось, что МД-моделирование, основанное на расчетах сродства связывания с использованием молекулярной механики при использовании уравнения Пуассона–Больцмана, внесет значительный вклад в решение реальных проблем, таких как идентификация самых выгодных комбинаций для пар «белок–лиганд» с их последующей оптимизацией.

Глубокое обучение: новая глава в моделировании белок–лигандных взаимодействий

Первые методы глубокого обучения появились в середине 1960-х, однако обрели популярность только к середине 2000-х гг. с возрастанием вычислительных мощностей и появлением объемных наборов экспериментальных данных. В настоящее время применение технологий глубокого обучения к задачам в различных сферах дало результаты, не уступающие, а иногда и превосходящие результаты традиционных методов. Самый яркий пример — создание алгоритма AlphaFold-2, предсказывающего третичную структуру белка по первичной аминокислотной последовательности в течение всего лишь нескольких минут [8], что стало революцией в области структурной биологии.

Модели глубокого обучения были предложены для прогнозирования белок–лигандного взаимодействия в качестве альтернативы традиционному докингу, основанному на поиске минимума свободной энергии [9]. Преимущество глубокого обучения заключается в возможности изучать белок–лигандное взаимодействие непосредственно из пространственного расположения атомов, минуя выбор математических параметров, которые не всегда отражают реальный механизм связывания. Этот метод предсказания белок–лигандного взаимодействия в настоящее время претерпевает бурное развитие: опубликованная в 2017 г. нейросетевая модель DEEPsite [10] на определенном наборе данных корректно определила 23,8% сайтов связывания лигандов, а опубликованная в 2020 г. нейросеть Kalasanty на той же выборке показала результат 44,6%. PUResNet в 2021 г. значительно улучшила результаты предсказаний (53% успеха у PUResNet против 51% у Kalasanty) [11].

К данному моменту разработано множество разнообразных алгоритмов работы и обучения нейронных сетей (рис. 2). При исследовании белок–лигандных взаимодействий обычно применяют сверточные нейронные сети (CNN), графовые нейронные сети (GNN) и сети-трансформеры. Сверточные нейросети рассматривают парные взаимоотношения между атомами через их взаимное расположение в пространстве. Принцип работы графовых нейросетей основан на учете пороговых значений для определения типа взаимодействия между атомами (ковалентное или нековалентное). Потенциальным преимуществом такого подхода является использование меньшего количества параметров. Используют также комбинации нескольких алгоритмов или добавляют иные модули (например, denoising autoencoder, удаляющий шум), которые улучшают конечный результат [12].

ЗАКЛЮЧЕНИЕ

Среди разных архитектур нейросетей пока не выявлен однозначный лидер: точность результата каждого алгоритма зависит от типа белков и лигандов, аффинности и механизма связывания. Согласно последним данным, графические нейросети egGNN и saCNN оказались наиболее успешными в предсказании аффинности лигандов [13, 14], однако отличие от сверточных нейросетей не является критическим. По нашему мнению, это связано с тем, что оптимальный предсказательный алгоритм пока не разработан. Создание такого алгоритма, судя по скорости развития ИИ в вычислительной биологии, — это, скорее всего, вопрос нескольких лет, а не десятилетий. Так же, как AlphaFold-2 в 2020 г. изменил парадигму в области изучения структуры белка, так и искусственный интеллект в биомедицинских исследованиях открывает новую главу в фармацевтике и драг-дизайне.

Перспективность использования ИИ для поиска лекарственных препаратов уже стала очевидна для индустрии, поскольку его применение существенно ускоряет и удешевляет классический двенадцатилетний цикл разработки нового лекарства. За последние пять лет почти все крупные фармкомпании объявили о партнерстве с ведущими ИИ-компаниями (Sanofi — Aily Labs, Pfizer — IBM, Novartis — Microsoft, AstraZeneca —

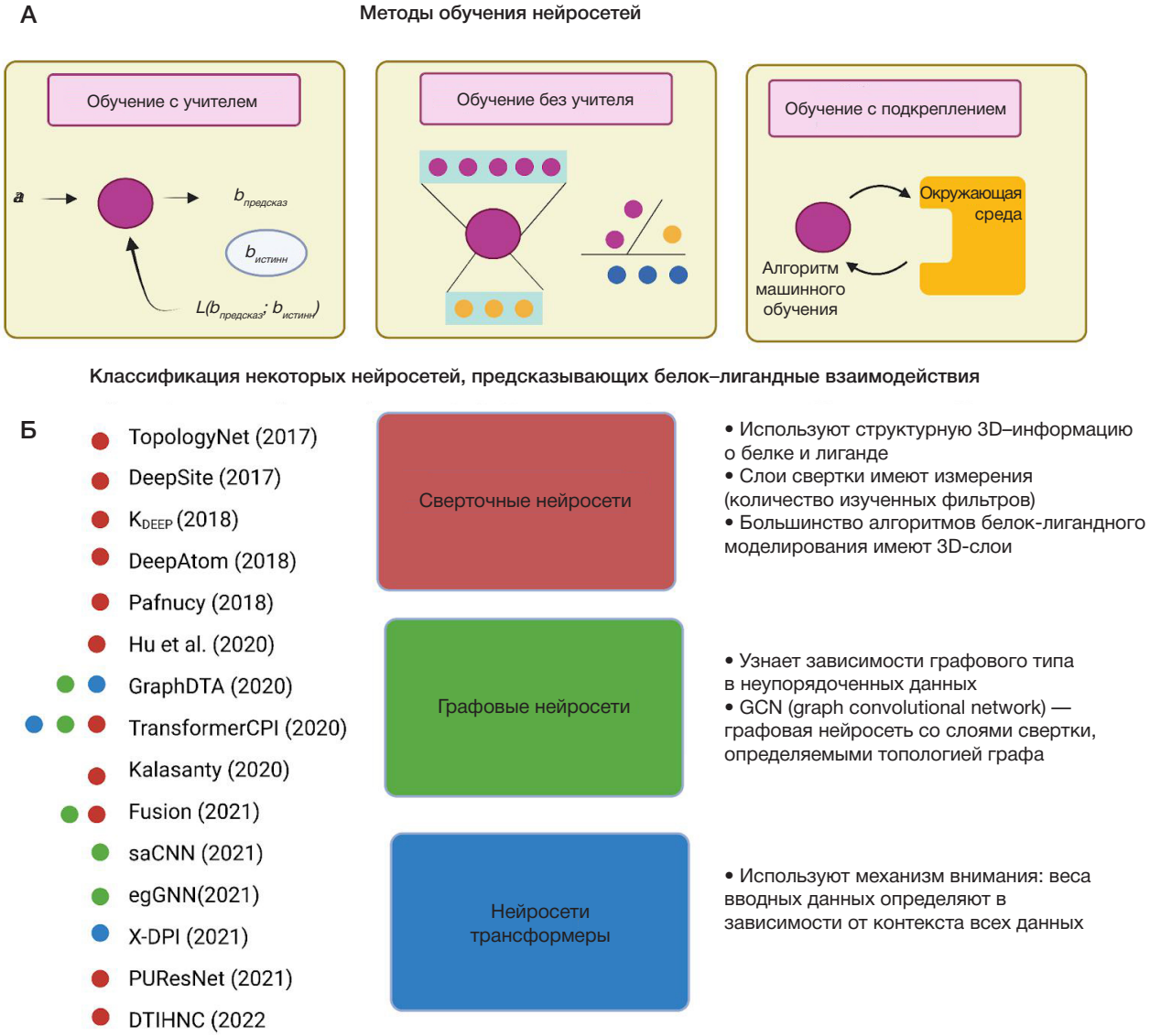

Рис. 2. Алгоритмы работы нейросетей. **А**. Методы обучения нейросетей. **Б**. Основные виды нейросетей, предсказывающих белок–лигандные взаимодействия

Benevolent и др.). Детали исследований, вероятно, еще долго будут защищены коммерческой тайной, однако регулярно выходят пресс-релизы, свидетельствующие о внедрении глубокого обучения в текущие R&D-процессы. Кроме того, появляется все больше сообщений об успехе лекарственных кандидатов, разработанных с помощью ИИ, которые готовятся к прохождению или находятся в клинических испытаниях. Примерами таковых являются халицин (перспективный антибиотик широкого действия, доклинические исследования) [15], INS018_055 (препарат от идиопатического легочного фиброза, 2-я стадия клинических испытаний), REC-2282, REC-994, REC-4881, BEN-2293, EXS-21546, RLY-4008, EXS-4318, BEN-8744 и др. [16].

Мы считаем, что в ближайшем будущем поиск пула новых биоактивных молекул кардинально ускорится, а глубокое обучение станет обязательной частью процесса разработки новых лекарственных препаратов. Тем не менее, до сих пор одной из актуальных проблем на пути к повсеместному применению глубокого обучения при поиске лекарств остается грамотный подбор данных, используемых для обучения нейросетевой модели, поскольку от их качества критически зависит достоверность ее предсказаний. В связи с этим задача максимально эффективного обучения на неполных или небольших наборах данных остается главным вызовом перед ИИ в сфере разработки лекарств [17].

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.

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.