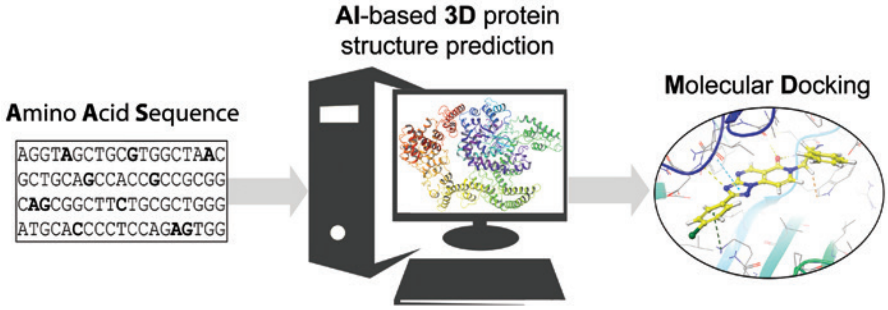
https://doi.org/10.59761/RCR5107

AlphaFold for a medicinal chemist: tool or toy?

Yan A. Ivanenkov,[a,b] Sergei A. Evteev,[a,b] Alexander S. Malyshev,[a,b] Victor A. Terentiev,[a,b] Dmitry S. Bezrukov,[c] Alexey V. Ereshchenko,[a,b] Anastasia A. Korzhenevskaya,[a] Bogdan A. Zagribelnyy,[c] Petr V. Shegai,[a] Andrey D. Kaprin[a,d]

[a] *P.Hertsen Moscow Oncology Research Institute,
2nd Botkinsky proezd 3, 125284 Moscow, Russian Federation*
[b] *Dukhov Automatics Research Institute (VNIIA),
ul. Sushchevskaya 22, 127030 Moscow, Russian Federation*
[c] *Department of Chemistry, Lomonosov Moscow State University,
Leninskie Gory 1, stroenie 3, 119991 Moscow, Russian Federation*
[d] *Peoples' Friendship University of Russia (RUDN),
ul. Miklukho-Maklaya 6, 117198 Moscow, Russian Federation*



The development of novel small drug molecules is a complex and important cross-disciplinary task. In the early stages of development, chemoinformatics and bioinformatics methods are routinely used to reduce the cost of finding a lead compound. Among the tools of medicinal chemistry, docking and molecular dynamics occupy a special place. These methods are used to predict the possible mechanism of binding of a potential ligand to a protein target. However, in order to perform a docking study, it is necessary to know the spatial structure of the protein under investigation. Although databases of crystallographic structures are available, the three-dimensional representations of many protein molecules have not been reported. There is therefore a need to model such three-dimensional conformations. Several computer algorithms have been published to solve this problem. AlphaFold is considered by the scientific community to be the most effective approach to predicting the three-dimensional structure of proteins. However, the scope of its application in medicinal chemistry, especially for virtual screening, remains unclear. This review describes methods for predicting the three-dimensional structure of a protein and provides representative examples of the use of AlphaFold for the design and rational selection of potential ligands. Special attention is given to publications presenting the results of experimental validation of the approach. On the basis of performed analysis, the main problems in the field and possible ways to solve them are formulated.
The bibliography includes 154 references.

Keywords: AlphaFold, medicinal chemistry, docking, drug, neural networks, machine learning.

Contents

1. Introduction

The development of new small drug molecules is a high-tech, phased, and expensive process.[1] To reduce the time and financial expenses in the early stages of development, various computational modelling methods[2,3] are typically employed. This allows for the analysis of, for example, the probable mechanism of ligand binding to a target, selection of the most promising molecules for high-throughput virtual screening (VS)[4] stage, prediction of their pharmacokinetic parameters,[5] suggestion of retrosynthetic pathways,[6] and evaluation of other important properties. Historically, the following key approaches to virtual screening have formed:[7]

— ligand-based drug design, LBDD;
— structure-based drug design, SBDD.

Both approaches have their strengths and weaknesses; however, the latter method can be considered more accurate and informative. It is evident that to carry out typical SBDD

Received 2 November 2023

(structure-based drug design) procedures, one must have access to the three-dimensional structure of the target (in most cases, this is a protein molecule; see Section 7 for the full names of the proteins reviewed in this overview). Such information can be found in specialized sources, such as the PDB[a] database, which contains data on the three-dimensional structures of proteins.[8] However, for many targets, especially new ones, such information may be lacking. In these cases, the method of constructing three-dimensional structures by homology is often employed[9] if crystallographic data of comparatively high resolution (preferably not more than 3 Å) are available for a closely related template. It should be noted that until recently, there were no available effective and accurate methods for solving such a task within the scope of VS.

The rapid development of machine learning and artificial intelligence methods has led to the creation of computational algorithms that, among other things, allow for the prediction of three-dimensional protein structures.[10–12] This undoubtedly holds great significance for bioinformatics, structural biology, enzymology, in particular for the study of enzyme action mechanisms and self-regulation,[13] the investigation of oncogenic mutations,[14] modelling of protein–protein interactions,[15] *etc.* Such methods play an important role in X-ray crystallography[16,17] and in conducting broad statistical studies of various functional and structural properties. Overall, information about their three-dimensional geometry is necessary for the creation of protein molecules; this information also contributes to a deeper understanding of protein functions under normal physiological conditions and in pathologies.

Despite the obvious successes in predicting the three-dimensional structure of proteins, the effectiveness of such an approach depends on the specific task. In medicinal chemistry, significant attention is paid to the mechanism of interaction between the ligand (small molecule) and the binding site through the formation of chemical bonds between the atoms of the ligand and the amino acids that constitute the binding site, which largely determines its affinity and selectivity. Furthermore, it is necessary to consider the conformational flexibility and heterogeneity of the sites,[18] the nature of the solvent molecules[19,20] filling the site cavity, including bridge water molecules, and in some cases, the critical role of just one or two amino acids that determine the activity and selectivity among the closest homologues in the family, for example, as in the case of CDK5/2 kinases (Asn144/Asp144),[21] Wee1/2 (Asp386/Ala386).[22] The solvent effect for polar amino acids forming the active site of the enzyme is generally weaker. The spatial position of such ensembles is deterministic, which is reflected in their B-factor values. The relatively low conformational mobility of amino acids is ensured, among other things, by hydrogen bonds. However, differences in substrate specificity are observed even among close homologues, despite their considerable overall spatial similarity. For example, among ATP-competitive kinase inhibitors (ATP stands for adenosine triphosphate), selectivity is determined by interaction both with amino acids in the hinge region and with distant pockets, which, in particular, defines the type of kinase inhibitors.[23] In other words, in many protein molecules, depending on their functions, there are conservative areas the spatial geometry and mutual arrangement of which can be predicted with relatively high accuracy by modern methods. Examples include evolutionarily developed domain structures and regions characterized by relatively high

[a] https://www.rcsb.org/ (access 28.03.2024).

Ya.A.Ivanenkov. Head of the Laboratory of Medical Chemistry and Computational Modeling (Moscow Research Oncology Institute named after P.A.Herzen). Head of the Laboratory of Medical Chemistry and Chemoinformatics (VNIIA).
E-mail: yaiivanenkov@gmail.com
Current research interests: development of novel small-molecule drugs, medicinal chemistry, organic chemistry, biochemistry, machine-learning algorithms and AI in the field of drug design and development, QSAR/QSPR analysis, pharmacokinetics and pharmacodynamics, hit-to-lead (H2L)/lead-to-drug (L2D) optimization, virtual screening, chemoinformatics.
S.A.Evteev. Leading scientist. E-mail: evteevsa1320@gmail.com
Current research interests: medicinal chemistry, chemoinformatics, molecular modeling.
A.S.Malyshev. Researcher. E-mail: alexmalyshev95@gmail.com
Current research interests: medicinal chemistry, chemoinformatics, molecular modeling.
V.A.Terentiev. Researcher. E-mail: terentiev.victor1@gmail.com
Current research interests: molecular modeling, medicinal chemistry, chemoinformatics, neural networks.
D.S.Bezrukov. PhD, associate professor at the Department of Chemistry, MSU. E-mail: bezrukovds@my.msu.ru
Current research interests: mathematical and quantum chemistry, molecular modeling.
A.V.Ereshchenko. Researcher.
E-mail: ereshchenko.alexey@gmail.com
Current research interests: artificial intelligence, machine learning, data science.
A.A.Korzhenevskaya. Researcher, 4th year student at High School of Economics (HSE), Faculty of Chemistry.
E-mail: aakorzhenevskaya@gmail.com
Current research interests: medicinal chemistry, organic chemistry, chemoinformatics, neural nets, development of small-molecule drug compounds.
B.A.Zagribelnyy. PhD student, Division of Medicinal Chemistry and Fine Organic Synthesis, Department of Chemistry of the MSU.
E-mail: bogdan.zagribelnyi@chemistry.msu.ru
Current research interests: generative chemistry, medicinal chemistry, retrosynthesis, small molecules design.
P.V.Shegai. Deputy General Director for Science of the Institution 'National Medical Research Center of Radiology' of the Russian Ministry of Health. Head of the Center for Innovative Radiological and Regenerative Technologies.
E-mail: contact.shegai@mail.ru
Current research interests: diagnosis and treatment of oncological and urological diseases, development of molecular means of targeted delivery of diagnostic and therapeutic anticancer drugs, radio- and chemotherapy.
A.D.Kaprin. Doctor of Medical Sciences, Professor, Academician of the Russian Academy of Sciences, Academician of the Russian Academy of Education. General Director of the Institution 'National Medical Research Center of Radiology' of the Ministry of Health of Russia. President of the Association of Oncologists of Russia. President of the Association of Directors of Centers and Institutes of Oncology and X-ray Radiology of the CIS and Eurasian Countries. Honored Doctor of the Russian Federation. Twice Laureate of the Government Prize of the Russian Federation in the field of science and technology. Chief oncologist of the Ministry of Health of the Russian Federation. Head of the Department of Oncology and X-ray Radiology named after Academician V.P.Kharchenko of the MI RUDN.
E-mail: kaprin@mail.ru
Current research interests: diagnosis and treatment of oncological and urological diseases, creation of new areas of training and development of educational trajectory for specialists in the field of nuclear medicine.

mobility, which often influences the binding of a ligand to the protein target (protein – ligand docking).

Molecular docking methods play an important role in the arsenal of the modern medicinal chemist. They are used to assess the possibility of ligand binding to a specific area in the target structure. It is clear that in modelling scenarios where the amino acids lining the binding site remain stationary and the three-dimensional geometry of the ligand is optimized to minimize the scoring function, even seemingly minor differences in the pocket geometry can lead to invalid results from the computational experiment. When using flexible docking, which optimizes the positions of the ligand atoms and surrounding amino acids simultaneously, the effect of such differences is mitigated, but assessing the result remains challenging, as in most cases only one 'trajectory' along the gradient of potential energy is selected. A more detailed mechanism of ligand interaction with the target can be obtained through a dynamic description of the binding process using molecular dynamics (MD) methods.[24,25] However, such approaches involve significant time expenditures and require substantial computational power. It should be noted that not all ligands bind according to the key – lock principle; many interact with the binding site according to the induced fit mechanism, which is associated with conformational transitions in the pocket structure. Similarly, drugs acting on allosteric sites lead to changes in the geometry of the binding site with the endogenous ligand, unless the discussion involves the regulation of protein–protein interactions. Given the above, the question of the applicability limits of neural network algorithms, which predict the three-dimensional structure of proteins for the needs of medicinal chemists, especially in the early stages of development, remains open.

2. Computer algorithms for predicting the three-dimensional structure of protein molecules

Anfinsen's seminal study[26] conducted in the 1970s demonstrated that the tertiary structure of a protein depends on its amino acid composition. The development of sequencing methods has allowed for the identification of more than 2.6 billion nucleotide sequences as of 2021,[b] of which more than 200 million have been translated into corresponding amino acid sequences.[c] However, the amino acid composition alone provides only a limited understanding of the biological functions of a protein, since these functions are determined by its spatial structure. For instance, nearly 30% of the human proteome consists of disordered structures that perform their functions despite the lack of well-defined tertiary structures.[27] Specifically, the active conformation of such proteins may form at the moment of interaction with a partner molecule.[28] Owing to crystallography and NMR spectroscopy methods, there are currently more than 200,000 three-dimensional conformations of protein molecules in the protein structure database, which is less than 0.1% of the total number of sequences in the UniProt database. Over time, the gap between the number of known protein sequences and experimentally determined tertiary structures is narrowing. In recent years, the spatial structures of many proteins have been modelled using various computational methods. The development of high-precision protein structure prediction methodologies is the most promising approach to bridging the discrepancy between the number of known amino acid sequences and the number of experimentally determined three-dimensional protein conformations.

Research aimed at predicting the tertiary structure of protein molecules began in the mid-20th century when the first Ramachandran plots[29] were published and methods for constructing three-dimensional structures based on amino acid sequence homology were proposed.[30,31] Later, the first software emerged, including the MODELLER and SWISS-MODEL algorithms to model the protein tertiary structures based on homology. For instance, for proteins with amino acid homology over 50% relative to the template, the average root mean square deviation (RMSD) may not exceed 1 Å compared to the experimentally determined structure. Alignment is typically carried out using the Needleman – Wunsch or Smith – Waterman algorithm,[32,33] with relatively fast alignment achievable using programs like BLAST,[34] PSSMs and HMMs,[35,36] 3D-Jury[37] and LOMETS.[38] It is clear that as homology decreases, the deviation increases, averaging 2 – 5 Å for homology around 30 – 50%, while for proteins with <30% homology, modelling results are mostly considered unreliable.[39,40] Identifying remotely homologous templates is no less challenging,[41] reducing the likelihood of selecting the most suitable structure for modelling.

Currently, various neural network algorithms for the modelling of three-dimensional structures of protein molecules are available. The general goal of such approaches is to predict the spatial position of each atom in the protein molecule based on its amino acid sequence. Some methods use available templates for modelling, while others do not require the presence of homologous three-dimensional structures. During the scientific competition CASP-14 (CASP is Critical Assessment of Protein Structure Prediction), it was shown that using an end-to-end approach based on deep learning algorithms,[42] it is possible to predict the coordinates of amino acid atoms in single-domain protein molecules with high accuracy, in particular without the direct use of templates. The quality of the prediction was weakly correlated with the number of available homologues. The most well-known algorithms that allow modelling of the tertiary structure of a protein based on its amino acid sequence are presented in Table 1. As an example, let us take a closer look at some of them.

One of the successful strategies for modelling the three-dimensional structures of protein molecules with distant homology is I-TASSER.[56] For template searching, the algorithm uses the threading method, performed by the LOMETS (local meta-threading server)[38] module. In the structure of the identified templates, closely related fragments and their three-dimensional conformations corresponding to sections of the studied amino acid sequence are determined. These fragments are then assembled into a set of tertiary conformations of a 'virtual' protein, which has the maximum similarity in relation to the studied amino acid sequence. This stage is carried out using the Monte Carlo method and the DOOP (docking decoy-based optimized potential) algorithm,[57] which uses empirically distance-dependent interactions between pairs of atoms and considers elements of the secondary structure. As a result, thousands of possible three-dimensional structures are generated. Next, a clustering procedure based on structural similarity is carried out. Statistical processing of the clustering results allows for the identification of the most probable three-dimensional structures from the formed clusters, with the scoring function being calculated using SPICKER.[58] Then, a reassembly

[b] GenBank. https://www.ncbi.nlm.nih.gov/genbank/ (access 28.03.2024).

[c] UniProt. https://www.uniprot.org/ (access 28.03.2024).

Table 1. Examples of software for predicting the three-dimensional structure of protein molecules.

Algorithm name	First publication year	Short description	Ref.
Crystallography & NMR System (CNS)	1997	Predicting the three-dimensional structure of a protein is based on experimental crystallographic data or NMR	43
ROSETTA	2001	For predicting three-dimensional conformations of proteins, the Monte Carlo algorithm and a proprietary energy scoring function are used	44
HHpred	2005	The method is based on searching for homologous templates using a hidden Markov model, particularly searching for examples with distant homology. The MODELLER algorithm is used to create three-dimensional models	45, 46
Pcons	2001	The algorithm compiles predicted three-dimensional protein structures from various sources to create a consolidated final forecast, enhancing the accuracy of the resulting structure	47
I-TASSER (iterative threading assembly refinement)	2008	An iterative approach that utilizes structural templates from known databases to predict the three-dimensional structure of protein molecules and employs a fragment method	48
Phyre2	2011	The method is based on detecting homology between related proteins with known structures; taking these data into account, the structure of the protein of interest is predicted	49
QUARK	2012	The *de novo* method is based on constructing the protein structure from fragments, without using information about known structural templates	50
AlphaFold-1	2018	The method is based on deep learning and predicts three-dimensional structures of protein molecules. The algorithm uses the ResNet architecture for modelling the distance map, as well as the values of the dihedral angles ψ and φ	51
AlphaFold-2	2020	An improved version of AlphaFold-1, which uses the Transformer architecture to model the relationships between amino acids	52
SPARKS-X (sequence, secondary structure profiles and residue-level knowledge-based energy score)	2020	The method is based on the alignment of primary sequences and structures of proteins: it combines *ab initio* protein folding and the prediction of three-dimensional structure based on templates, considering the implicit influence of solvent molecules	36
OmegaFold	2022	Analogous to AlphaFold, however, the architecture and size of neural networks differ. The template search is not conducted by classical methods but through a neural network approach	53
ESMFold	2023	The method utilizes a language neural model (Transformer ESM-2), surpassing AlphaFold in speed but inferior in accuracy. The program predicted the structures of 600 million proteins within two weeks	54
SWISS-MODEL	2009	The construction of a three-dimensional protein model includes four main stages: searching for a suitable structural template, aligning the study sequence with the template, building the three-dimensional model, and assessing the quality of the model	55

procedure is carried out using the LOMETS, TM-align (Ref. 59), and IRP (inherent reduced potential) algorithms. The final full-atom three-dimensional protein model is created during optimization by the modified REMO H-Bond algorithm[60] using a set of empirical rules. Unaligned sections are also minimized. A key feature of I-TASSER is the use of sets of structural templates. Examples of using this algorithm to solve various tasks in the field of bioinformatics and the development of new drug molecules are described in a number of publications (see, for example, Refs 61–65).

Before the emergence of AlphaFold, the ROSETTA algorithm[44, 66] was leading in the CASP (Critical Assessment of Protein Structure Prediction) competitions. The program is based on the fragment insertion approach, which uses relatively short segments from known protein structures to initiate modelling. For each small fragment (3–9 amino acids) of the protein sequence, the algorithm searches for suitable three-dimensional templates and randomly selects them as starting representations with a homology threshold of 50%. Alignment is performed using the PSIBLAST algorithm,[67] involving full pairwise comparison. Fragments are described by torsion angles from a library of training examples (crystallography data with a resolution of no more than 2.5 Å). For predicting secondary structures, ROSETTA uses methods like Psipred, SAM-T99, and JUFO; torsion angles are compared with corresponding angles from the training examples, and erroneous results are excluded from consideration. Next, assembly and optimization procedures of the initial three-dimensional conformations are carried out using the Monte Carlo method and an energy function. At this stage, the algorithm introduces random changes to the values of torsion angles and ranks changes with a certain probability, according to the Metropolis criterion.[68] ROSETTA uses a specialized empirical function for calculating potential energy to evaluate the results, taking into account the Lennard-Jones potential, solvation, and intermolecular hydrogen bonds. The goal of the optimization stage is to select the most stable conformations for a given amino acid sequence. A feature of the program is that the algorithm does not just create one structure and evaluate it but generates multiple possible variants and ranks them based on the scoring function. In each iteration, ROSETTA refines and improves the prediction outcome.

The I-TASSER and ROSETTA approaches are generally similar; however, the latter does not use the threading procedure and employs its own empirical function to estimate potential energy using the Metropolis criterion. ROSETTA uses a multitude of different starting structure variants, and the databases of initial templates also differ. The program operates in multiple threads, resulting in a large set of structures, with close conformations being clustered together. Sparse groups are not considered, and from the remaining ones, those with the lowest average energy are selected. In the first stage, ROSETTA essentially annotates the studied amino acid sequence for the potential presence of conservative secondary conformations, comparing small fragments of the sample with a reference database. This procedure is probabilistic in nature, as it involves many independent runs. Each amino acid is represented by a limited set of anchor points (practically, harmacophore centres) to accelerate the modelling process in the early stages. As a result, the program presents several variants of possible packings with minimal potential energy values. A full-atom model is then constructed and undergoes final optimization. However, the program does not perform best with proteins longer than 150 amino acids and requires substantial computational power and typically significant time expenditures. Examples of using ROSETTA algorithm can be found in a number of publications.[69–72]

SPARKS-X is the latest version of the program for sequence, secondary structure profiles, and residue-level knowledge-based energy score, which predicts the three-dimensional structure of a protein based on its amino acid sequence (previous versions include SPARKS, SP2, SP3, SP4, and SP5).[36] The algorithm is based on the method of multiple sequence alignment,[67] secondary structure prediction, its own scoring function[73] (modified SKSP method),[74] comparison method using template structures,[75] evaluation of the solvent accessible surface area (SASA),[76] and a torsion angle prediction module[77] using machine learning methods. In the CASP-6 and CASP-7 competitions, the SPARKS, SP3, and SP4 programs were among those that showed the best results.[78,79] Unlike previous versions, SPARKS-X integrates an updated energy function using hidden Markov models.[80] This improvement has enhanced the prediction quality for secondary structure based on the primary sequence: $Q^3 = 81-82\%$ (the proportion of correctly predicted substructures; Q^3 is a calculated parameter reflecting the accuracy of predicting a protein secondary structure: α helices, β-sheets, and loops[81]). The mean absolute error values for ψ and φ angles were 33° and 22°, respectively, with a determination coefficient $R^2 = 0.74$ for SASA. The alignment procedure on template sequences is implemented using a modified function from SP5 and the PSIBLAST method, with the optimization of the scoring function being performed according to the Smith–Waterman algorithm,[33] and template structures ranked according to the value of the standard statistical Z-score. Specifically, the developers managed to improve the overlay of C_α atoms according to the obtained values of the scoring function MaxSub (Ref. 82) for different protein families.

Protein secondary structure prediction involves several stages. First, a PSSM (position specific scoring matrix) is constructed using the PSIBLAST mutational profile and seven parameters reflecting steric hindrance, hydrophobicity, volume, polarizability, isoelectric point, and the probability of forming α-helices and β-sheets.[83] These parameters, along with the PSSM, serve as the input vector to a neural network. The secondary structure is evaluated using the composite SKSP7 scoring function.[84] In the next step, another neural network with a RealSPINE architecture[85] predicts the SASA (solvent accessible surface area) for amino acid residues, using the PSSM, the above parameters, and the predicted secondary structure as input vectors. Then, the SASA values, secondary structure, PSSM, and parameters are used to predict torsion angles (τ_0). Subsequently, the secondary structure for helical segments and β-sheets, to which incorrect torsion angles were initially assigned, is predicted using SASA, τ_0 values obtained in the second stage, PSSM, and parameters (refinement procedure). The newly obtained secondary structures, along with SASA, PSSM, and the specified parameters, are used to predict new torsion angles (τ_1). In the final stage, a neural network is used, the predictive ability of which was assessed using examples from the DSSP database. The authors chose PSSM, parameters, SASA, and τ_1 as the feature input vector. Neural networks were trained using a reference set of structures consisting of 2640 protein molecules obtained from the PISCES database[86] with homology not exceeding 25% (crystal resolution less than 3 Å), excluding molecules with a chain length exceeding 500 residues. Using test examples from CASP-9, it was demonstrated that the SPARKS-X algorithm can model three-dimensional structures of protein molecules, with the quality of modelling being comparable to that of ROSETTA. Results of the SPARKS-X algorithm can be found in a number of publications (see, for example, Refs 87–89).

3. AlphaFold algorithm

The AlphaFold algorithm has recently gained widespread recognition and attracted the attention of many specialists in the field of structural biology and bioinformatics. In 2023, the developers of this program, D.Hassabis and J.Jumper, were awarded the Lasker Award in the category of Basic Medical Research. The AlphaFold program, built on artificial intelligence, was trained on a large number of examples. Currently, two versions of the program are available. In December 2018, AlphaFold-1 ranked first in the overall CASP-13 competition. The program successfully tackled the task of predicting the spatial structures of proteins for which no template structures were available, despite partially similar amino acid sequences. In 2022, AlphaFold-2 won the CASP-14 competition: the program was able to relatively accurately model the three-dimensional structure of 35% of proteins for which close templates were absent, and for 77% of proteins with available templates.[52] The new version demonstrated higher accuracy compared to other similar algorithms and scored more than 90 out of 100 possible points for two-thirds of the proteins in the GDT (global distance test), which is used to assess how accurately a three-dimensional protein structure has been predicted compared to experimental data. For 88 out of 97 proteins, the AlphaFold-2 algorithm showed better results than other methods. In 2021, an article[52] was published (with more than 15000 citations as of October 2023) describing the AlphaFold-2 algorithm along with open-source software and the corresponding database.

In the first stage of the AlphaFold-1 algorithm, the sequence of the protein of interest is aligned with the sequences of other proteins with known three-dimensional structures (multiple alignment). Next, using the coevolution matrix, pairs of amino acids forming key (conservative) interactions are analyzed. A high correlation corresponds to critical contacts and indicates that the amino acids are close to each other in three-dimensional space. In homologous proteins, such pairs remain unchanged or change synchronously. Unlike other programs, AlphaFold-1 predicts pairwise distances between C_β atoms, providing discrete probability distribution densities for each amino acid in the form of matrices. A distance not exceeding 8 Å is used as the interaction threshold. The three-dimensional structure is reconstructed using a potential (analogous to a potential energy scoring function) dependent on the dihedral angles ψ and φ, pairwise positioning of C_β atoms, and volume overlaps. Optimization of this function is achieved through gradient descent. The locations of C_α atoms are mechanistically determined from known $C-C_\alpha-N$ distances. Modelling is performed using a ResNet class convolutional neural network[90] with a training sample of over 30 000 crystallographic structures. A key feature of AlphaFold is the structure of its input and output data.

Unlike AlphaFold-1, the end-to-end model AlphaFold-2 uses a transformer architecture with an ‘attention mechanism’. Training was conducted on 170 000 examples, which provided more accurate predictions in the CASP-14 framework and sped up the algorithm’s operation. For searching the closest homologues and aligning amino acid sequences, the program uses the HMMER method (hidden Markov models),[91] based on Markov models (chains), and databases such as Uniprot and MGnify, while the HH-suite method[92] is used to search for the closest three-dimensional structures in the PDB database. Like AlphaFold-1, the new version operates with matrices of pairwise distances, but in the absence of a suitable template, it is filled with default values. The neural network receives a low-dimensional continuous vector representation of the alignment and pairwise correspondence (vector representation) as input and iteratively improves their quality on deeper layers. As a result, the model returns the predicted three-dimensional structure of the protein.

AlphaFold-2 essentially consists of three modules: Evoformer (vector representation, computational block with trainable weights), the structural module (generation of three-dimensional structure), and OpenMM (optimization of atom coordinates–system relaxation). The main neural network of AlphaFold-2 predicts the positioning of the peptide backbone, while Deep ResNet outputs dihedral angles that describe the positions of amino acid side chains. The Noisy Student Training approach (semi-supervised learning) is used for training.[93] In the initial stages of the system operation, all elements of the peptide backbone are placed at the origin, and then the coordinates are updated by the structural module, taking into account the information received from the IPA (invariant point attention) function. AlphaFold-2 uses the FAPE (frame aligned point error) as its loss function, which is independent of changes in the global coordinate system. In terms of operation and data processing logic, AlphaFold-2 is virtually indistinguishable from the first version, but the neural network architecture and the training algorithm itself have been significantly modified. This, in particular, considering the use of a large volume of training data and the coevolution matrix, has allowed AlphaFold-2 to take a leading position in the field.[94]

We did not intend to conduct a detailed comparison of algorithms and their neural network architectures for predicting the three-dimensional structure of protein molecules. A recent publication[95] presents the results of a comparative analysis of the AlphaFold-2 and RoseTTAFold algorithms and describes details of their architecture. For instance, it highlights the significantly higher efficiency of the two-track neural network architecture used in AlphaFold-2 compared to RoseTTAFold. Furthermore, as another distinguishing feature, the authors pointed out that the multiple alignment parameters are updated in AlphaFold based on pairwise characteristics through the direct attention mechanism, which provides a more accurate prediction of the spatial positioning of atoms. AlphaFold-2 employs end-to-end training, updating all model parameters through the backpropagation of errors from the loss function, which is calculated from the three-dimensional coordinates after passing through several SE(3)-equivariant layers of the transformer neural network. More details on the algorithm features can be found in the supplementary materials to the mentioned article.

Although there are quite a few publications to date with examples of using the AlphaFold-2 algorithm for solving bioinformatics tasks,[96] its application for the development of new small drug molecules, especially at the early stages, remains unexplored. To be successful, a modern medicinal chemist must have deep knowledge in areas such as organic synthesis, classical medicinal chemistry, and chemoinformatics. It was mentioned earlier that methods of molecular docking or MD (molecular dynamics) are often used to predict and understand the mechanisms of molecule binding to a chosen target, based on which a specialist can evaluate the potential of a small molecule and modify it to improve the affinity. This requires the three-dimensional structure of the protein molecule. Problematic are those cases when it comes to a protein for which no data on the three-dimensional structure are available and there are no closely related analogues. Considering that the researcher’s primary focus is precisely on the area of potential binding of a small molecule to the target, the question of how accurately the AlphaFold algorithm can predict the geometry of the binding site adapted for docking modelling remains open. Below, recent scientific publications are detailed, where the AlphaFold algorithm was applied with the aim of developing drug molecules and virtual screening.

4. AlphaFold exploitation for small drug molecule development

4.1. Virtual screening

Comparing the results of docking simulations obtained using the Glide module (see[d] and Ref. 97) with data from the DUD-E database (see[e] and Ref. 98) allowed for a preliminary assessment of the rationality and efficiency of using model structures available in AlphaFold for virtual screening (VS) in comparison with the use of crystallographic data. Within the scientific community, the DUD-E database, despite its shortcomings, is considered to be a standard set for conducting independent testing of computer models.[99] As the subject of their study, Zhang *et al.*[98] selected *holo* (in complex with a ligand) and *apo* (without a ligand) forms of proteins corresponding to the AlphaFold model and their optimized variants, which were obtained using the induced fit docking molecular dynamics (IFD-MD) protocol.[f] Two different forms (*holo* and *apo*) were chosen based on the premise that docking simulation results depend on the geometry and composition of the binding site, which in many cases differ for the mentioned forms.

IFD-MD is a separate module that combines traditional approaches and docking functions, pharmacophore analysis of the binding site, solvation assessment (WScore function),

[d] https://www.schrodinger.com/ (access 28.03.2024).

[e] http://dude.docking.org (access 28.03.2024).

[f] https://newsite.schrodinger.com/platform/products/ifd-md/ (access 28.03.2024).

sampling of amino acid side chains, and partially MD (molecular dynamics) methods to optimize the ligand position (modelling the induced fit mechanism) in its algorithm. The algorithm, which requires the presence of a template structure of the ligand–protein complex, was used, in particular, to obtain a greater number of *holo*-like structures based on *apo*-forms of AlphaFold protein models. It was noted that the MOE program[g] also implements similar functionality in the forced alignment mode. The authors[98] pointed out that their approach is applicable exclusively to protein models the binding site of which has the peptide backbone correctly positioned. The original AlphaFold models were compared with modified and optimized variants and with relevant experimental data. It was noted that docking in AlphaFold models for which experimental data on the three-dimensional structure of protein molecules are not available does not guarantee the discovery of hit molecules. It was shown that the application of the IFD-MD method leads to a significant improvement in docking results according to the BEDROC enrichment coefficient.[h] However, the cited publication[98] clarifies that all examples from the DUD-E database were already available in the PDB at the time of training the AlphaFold algorithm.

The DUD-E database contains structural information on 102 protein molecules belonging to various families, including kinases, nuclear receptors, proteases, and others. For each class of targets, a set of active ligands and inactive structurally similar molecules is available. For analysis, a set of 40 proteins was used, for which both forms (*holo* and *apo*) are presented in the database, by analogy with the study by Im *et al.*[100] Structures without analogues among the AlphaFold models for which only covalent ligands were indicated as targets were excluded from this set. After the filtering procedure, 27 proteins remained. Since for a number of objects, such as BRAF (B-Raf proto-oncogene), EGFR (epidermal growth factor receptor), IGF1R (insulin like growth factor 1 receptor), ITAL (integrin α-L), and RXRA (retinoic acid receptor-A), the binding sites of AlphaFold-2 (AF2) are blocked by an excess of amino acids, an alignment of the primary sequence between the AF2 model and the *holo*-form was performed, and then peptide segments before the first and after the last amino acid in the sequence corresponding to the *holo*-form were removed.

These operations resulted in a sample of modified AlphaFold models more suitable for reproducing the docking. Then, for all proteins and complexes, a preprocessing procedure was conducted using the ProteinPreparation module, during which hydrogen atoms were added, missing peptide chains were built, and protonation was carried out in accordance with the PROPKA (protein pK_a) algorithm at pH 7.4,[101] considering functionally significant tautomeric forms of histidine. Next, hydrogen bonds were optimized and a potential energy minimization procedure for the entire protein molecule was conducted. Active and inactive ligands from the DUD-E database underwent a standard preparation stage using the LigPrep module, with a maximum of 32 permissible starting conformations (default value). Cofactors and coenzymes that play a significant role in ligand binding and/or structural stabilization of the binding site were positioned in the corresponding AlphaFold models based on alignment, and their positions were refined during minimization. Using the IFD-MD protocol, models of complexes (no more than 5 different poses for each ligand) were obtained and used to perform docking simulation of active and inactive molecular structures. In some cases, the authors[98] used the PLDB module to select other reference molecules from the PDB database based on their structural similarity (Tanimoto metric, RDKit fingerprints) to the molecules associated with a specific target from the DUD-E set. Based on the results of molecular docking (Glide module, standard SP protocol, Table 2), one variant with the best scoring function was selected for each example. To compare the geometry of the binding sites (amino acids within no more than 5 Å from the ligand atoms were considered when calculating RMSD values) for *holo*- and *apo*-forms, as well as original and modified AlphaFold structures, the SiteMap module was used.

As a result, it was shown that the spatial geometry of the binding sites in AlphaFold structures resembles more closely the *holo*-forms (average $\text{RMSD}_{\text{backbones}}$ were 1.48 and 1.17 Å, and $\text{RMSD}_{\text{side chains}}$ were 2.15 and 1.83 Å for *apo*- and *holo*-forms, respectively). From the data presented in Table 2, it is evident that when using modified models, the docking results are comparable with those for the *apo*-forms of proteins but are significantly inferior to the results for *holo*-forms, which is explained by the principle of induced fit and protein-specific features, examples of which can include the DFG-*in/out* and αC helix *in/out* conformations of the kinase binding site. The applicability of such models for the needs of standard virtual screening is quite limited. Comparative target-specific data for the same sample but with the addition of IFD-MD-optimized structures are presented in Fig. 1.

Based on the results, it can be concluded that *holo*- and *apo*-forms with low RMSD values in the binding site area relative to the *holo*-forms yield the best docking results. A similar but less pronounced trend is observed for AlphaFold models. For example, in the RXRA protein molecule, the C-terminal helical segment is located inside the binding site, significantly affecting the docking results. Zhang *et al.*[98] noted that in the case of modified AlphaFold models, incorrect ligand positioning results were due in part to incorrectly defined rotamers of amino acids lining the site. Docking results for *holo*-and *apo*-forms, as well as for modified AlphaFold models compared to their IFD-MD-

Table 2. Docking results of small molecule structures from the DUD-E database with protein binding sites (experimental data for holo- and apo-forms, corresponding AlphaFold model with native and modified structure).[98]

Models	ROC	BEDROC (α = 160.9)	EF (1%) (see[†])	The proportion of active molecules based on docking results (%)
holo	0.804	0.407	22.4	97.3
apo	0.696	0.193	10.4	95.3
AlphaFold native	0.621	0.206	11.0	82.4
AlphaFold modified	0.664	0.199	11.0	90.8

[†] EF (enrichment factor) is calculated for a small portion of the database — the most highly scored compounds (1%).[102]

[g] https://www.chemcomp.com/Products.htm (access 28.03.2024).

[h] One of the most widely used tools for evaluating the performance of classification or ranking algorithms in statistics and machine learning are the receiver operating characteristic (ROC) curve and the area under the ROC curve (AuROC), which reflect the frequency of true positives and false positives. BEDROC is an ROC curve considering the Boltzmann distribution. When calculating BEDROC, the proportion of results that collectively provide 80% significance is set using the parameter α; this parameter also determines the requirements for the maximum allowable proportion of positive examples in the test at which the criterion does not reach saturation.

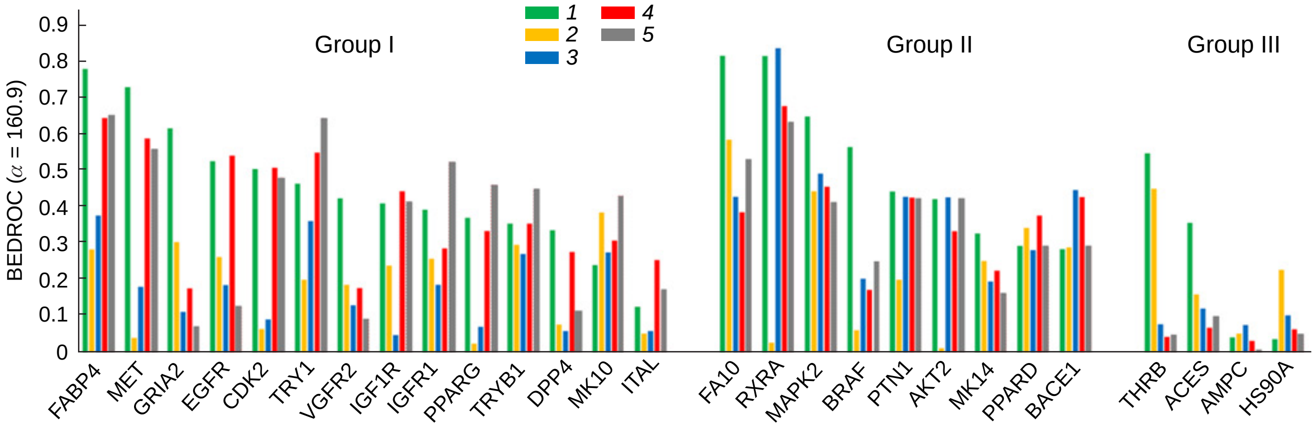


Figure 1. BEDROC enrichment coefficients for 27 targets. (*1*) *holo*-forms; (*2*) *apo*-forms; (*3*) modified AlphaFold models; (*4*) IFD-MD-optimized models, experimental ligand position data used for template structure construction; (*5*) IFD-MD-optimized models, active molecule pose used for template structure construction obtained during docking. Group I represents examples where models *4* showed better results compared to models *3*; Group II shows comparable results for models *1*, *3*, *4*; Group III consists of models *3* and *4* with low prediction quality compared to *holo*- and/or *apo*-forms.[98] The figure is published under CC BY-NC-ND 4.0 license.

optimized variants, in which reference ligands were preliminarily placed after the alignment or docking procedure, are presented in Table 3. However, it was emphasized that the poses obtained during docking did not always correspond to the true ones. IFD-MD-optimized models showed results comparable to those for *holo*-forms (see Table 3), especially in the cases where the entire set of docking models was used (see above). The authors effectively demonstrated the efficiency of using the IFD-MD method in relation to the EGFR kinase, for which *holo*- and *apo*-forms, as well as the modified AlphaFold model, were found to be unsuitable for docking construction. In particular, during optimization, different rotamers were obtained for the amino acid residues Met^{766} in the *C*-loop and Cys^{775}, freeing additional volume in the site for ligand binding, while the Glu^{762} side chain was displaced from the site. Conformational changes occurred to Phe^{856} in the DFG segment [Asp(D)–Phe(F)–Gly(G), a conservative segment in the structure of kinase pockets] and Thr^{790} (the gatekeeper region). Together, these spatial changes allowed the overall geometry of the binding site to be approximated to the *holo*-form and, consequently, improved the docking results. Similar experiments were conducted using reference structures of active molecules with low structural similarity to the ligands used previously, improving docking results, although this experiment involved significant time and computational expenses. Additionally, an approach was proposed that considers the influence of cofactors and coenzymes.

Table 3. Docking results for different models.[98]

Models	ROC	BEDROC (*α* = 160.9)	EF (1%)
holo	0.814	0.439	24.2
apo	0.733	0.213	11.4
AlphaFold modified	0.728	0.236	13.1
IFD-MD-AlphaFold optimized			
best pose, alignment	0.799	0.300	16.8
poses ensemble, docking	0.825	0.338	18.9
best pose, docking	0.783	0.281	15.6
poses ensemble, docking	0.817	0.325	18.0

Based on the experiments, Zhang *et al.*[98] concluded that unmodified AlphaFold structures yield results in standard docking modelling similar to those for *apo*-forms, which are noticeably inferior to the results for *holo*-variants. However, the use of the IFD-MD method significantly improves the quality of prediction. The authors did not conduct virtual screening using this algorithm to search for potential ligands, which does not fully assess the capabilities of the described approach, particularly in the field of developing new hit molecules acting on targets that do not have close homologues among available experimental complexes.

Weng *et al.*[103] used the AlphaFold model built for the protein WSB1 (WD repeat and SOCS box-containing protein-1) within the framework of virtual screening (VS) to search for inhibitors of its activity. This protein consists of seven WD40 domains (*N*-terminus) and one SOCS box (*C*-terminus). It promotes the growth and development of tumour cells by blocking the activity of the tumour suppressor protein pVHL and activating the transcription of the HIF-1α gene. The conducted virtual screening identified a number of compounds as potential ligands to the said protein, among which G490-0341 (**1**), G610-0188 (**2**), Y043-6168 (**3**), and Y044-5019 (**4**) showed the best results in subsequent MD studies (Fig. 2). Programs AutoDock-GPU (Ref. 104) and GlideSP (Schrödinger, Maestro[i]) were used to model the docking of structures.

In the first stage, the AlphaFold model was preprocessed using the MOE program, QuickPrep module, where structural correction, addition of hydrogen atoms, removal of unbound water molecules, and energy minimization were performed, while in the Maestro program, ligand structure preparation was carried out using the LigPrep module. The peptide ligand D2 (type 2 iodothyronine deiodinase) was used to identify the binding site using MD [binding pose meta dynamics (BPMD) method was applied]. The initial docking modelling was performed using the AutoDock-GPU program, with molecules from the ChemDiv collection[j] serving as the subjects of study. Based on the experiment results, more than 127 000 molecules with a scoring function not exceeding –10 kcal mol^{-1} were selected, which were then examined in a stepwise docking using

[i] https://www.schrodinger.com/products/glide (access 28.03.2024).
[j] https://www.chemdiv.com/ (access 28.03.2024).

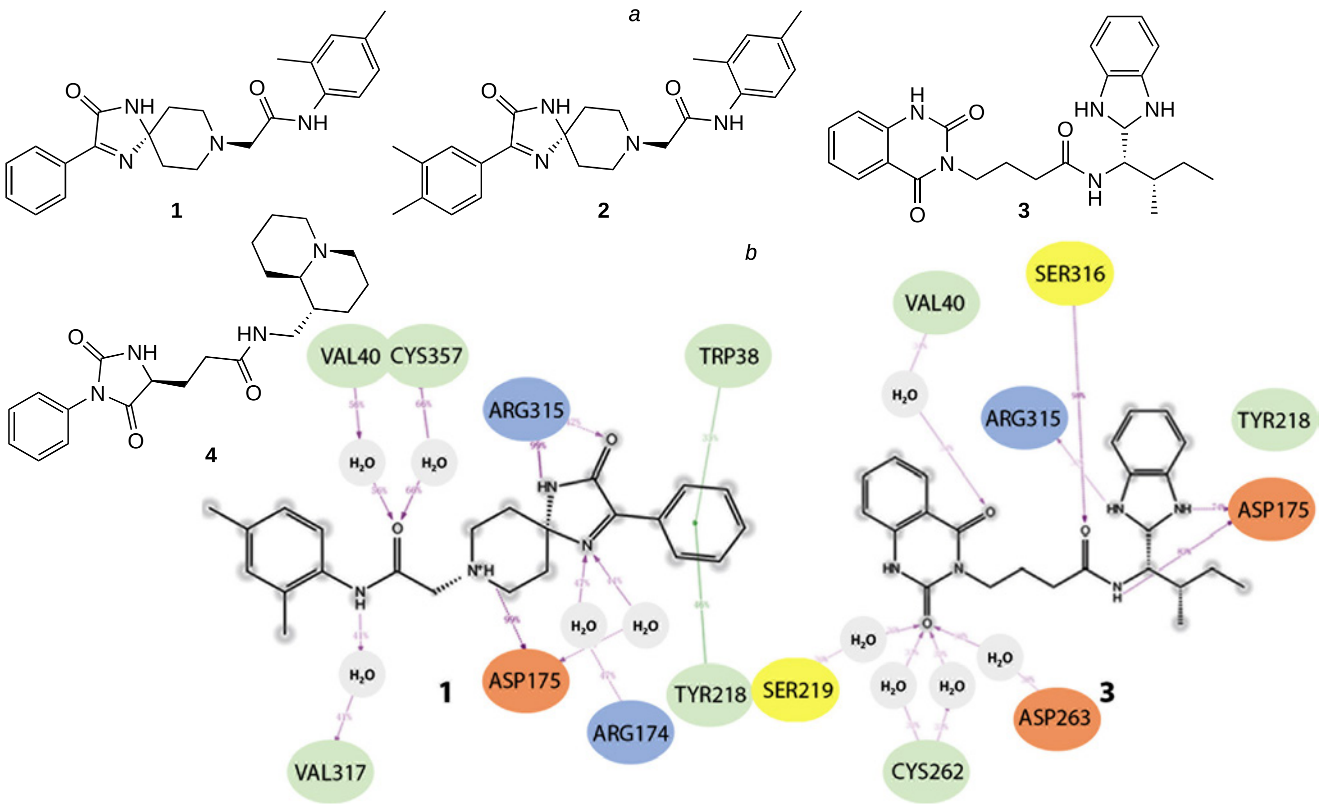


Figure 2. Structures of promising molecules, potential WSB1 inhibitors (*a*) and most likely binding mechanisms for molecules **1** and **3** based on MD results (*b*).[103] The figure is published under CC BY 4.0 DEED Attribution 4.0 International license.

the Maestro program. During the experiment, a maximum of 20 possible poses with a Glide XP score $\leqslant -8.0$ kcal mol^{-1} were obtained for each structure. Based on the modelling results, 20 most promising structures were selected, for which the BPMD method was also applied to refine their poses. It was shown that compound **1** is capable of forming strong bonds with key residues Asp^{175} and Arg^{315}, forming a stable complex. The authors [103] did not present the results of biological testing of the selected molecules, which does not fully assess the contribution of the AlphaFold model to the results of the described experiment.

Wong *et al.*[105] used AlphaFold structures and docking modelling to predict the possible interaction between 296 *E. coli* proteins and 218 antibacterial molecules. The approach was validated for 12 proteins *in vitro*. The study results showed an average auROC value of 0.48, indicating low efficiency of the approach. Reassessment of docking poses using machine learning methods improved the model characteristics (auROC = 0.63 – 0.71). The quality of prediction varied depending on the used activity threshold and averaged 41 – 73%. A negative sample of 100 inactive molecules was used. In the first stage, the authors investigated the antibacterial activity (inhibition of *E. coli* K-12 BW25113 growth) for a library consisting of more than 39 000 molecules, which included known antibiotics as well as natural compounds and synthetic molecules with high structural diversity (molecular weight from 40 to 4200 Da). Molecules were tested at a relatively high concentration (50 μM). Compounds that showed activity of at least 80% (218 molecules) were selected as hits, about 80% of which belonged to known classes of antibacterial agents, including β-lactams, aminoglycosides, tetracyclines, fluoroquinolones, *etc*. The docking procedure for the selected structures was carried out using AutoDockVina. As an independent *in silico* control, complexes available in the PDB database were used.

As a result, over 64 000 poses for active molecules and over 29,000 for inactive ones were predicted. At threshold scoring function values of –7 and –5 kcal mol^{-1}, respectively, 9.6 and 31% of molecules were classified as active against at least three proteins, while out of 296 proteins, 178 and 216 proteins showed a likelihood of binding to at least three active molecules. For inactive compounds, these indicators were respectively 86 and 99 molecules, and 137 and 204 proteins. Based on these results, it can be concluded that the model has a low classifying ability. In addition, the authors [105] assessed the predictive capabilities of the model using 142 antibiotic–target pairs described in the literature to replicate their binding mechanisms. However, unsatisfactory results were obtained in this case as well, regardless of the scoring function threshold. The quality of the model was studied based on *in vitro* testing results conducted for 12 *E. coli* proteins, including DNA gyrase, primase, helicase, NAD$^+$-dependent ligase, polymerase, guanylate kinase, mur family proteins, and others, for which test systems were available that allowed for the direct assessment of molecule binding to these proteins and their activity (Fig. 3).

In the biological experiment, it was found that 94 and 85 compounds at a concentration of 100 μM showed inhibitory capability exceeding 50% against murA and DNA helicase, respectively. Conversely, the number of molecules acting on DNA ligase and murC was much lower, at 4 and 5, respectively. For all primary hit molecules, IC$_{50}$ (half-maximal inhibitory concentration) values were experimentally determined. It turned out that 45 molecules showed nonspecific activity and could be classified as PAINS (pan-assay interference compounds). The

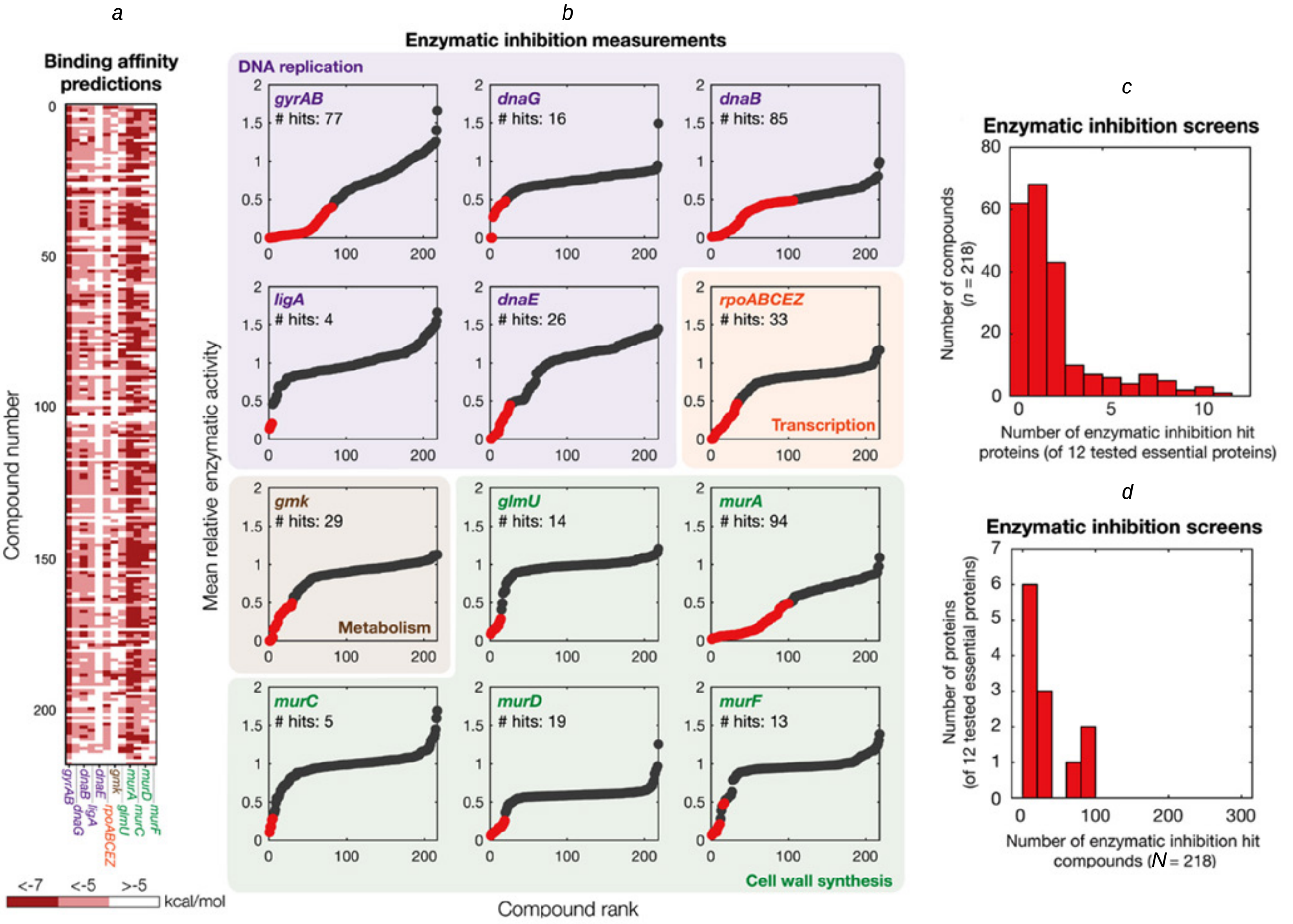


Figure 3. Predicted affinity matrix for 218 molecules and 12 bacterial proteins (*a*); results of primary biological testing (*b*); proportion of active molecules per single protein (*c*); proportion of proteins per single hit molecule (*d*).[105] The figure is published under CC BY 4.0 license.

proportions of correctly predicted hits were 59 and 30% with scoring function thresholds of –5 and –7 kcal/mol, respectively (Fig. 4*a*). For false-positive molecules, this indicator was 66 and 22% respectively (Fig. 4*b*). These results indicate that the model efficiency was not significantly different from the efficiency of a model with randomly selected molecules for screening. The auROC value for the 12 proteins ranged from 0.18 (murC) to 0.71 (gyrAB), with an average value not exceeding 0.5.

Similar results were obtained using auPRC (pose/ranking consensus). The authors[105] noted, in particular, that they observed no correlation between auROC (or auPRC) and pLDDT (predicted local distance difference test).[52] A similar computational experiment was conducted using molecules of the same compounds the structures of which are available in the PDB database. However, no improvements in the prediction quality were established. Consequently, Wong *et al.*[105] used another docking modelling method, namely DOCK6.9 (Ref. 106) and evaluated the influence of several scoring functions based on machine learning models (Figs 4*c*,*d*): RF-Score,[107] RF-Score-VS,[108] PLEC score,[109] and NNScore.[110] The results showed that the RF-Score, RF-Score-VS, and NNScore functions, applied to docking modelling results in AutoDockVina, significantly improved the average auROC values: 0.62, 0.63, and 0.58 respectively (across 12 proteins). Among the main conclusions, it was noted that AlphaFold-2 models, having several drawbacks, including the inability to distinguish between active and inactive protein conformations,[111] in conjunction with docking modelling results (ligand is mobile, protein atoms are not) in a high-throughput virtual screening mode, do not provide a prediction quality sufficient for rational selection of molecules for biological testing.

Scardino *et al.*[112] investigated the efficiency of high-throughput docking (HTD) using AlphaFold models and corresponding crystallographic structures of 22 proteins. The analysis provided the conclusion that the original AlphaFold models used for this purpose were of low efficiency. In particular, the authors noted the important role of proper model preparation before the virtual screening (VS) stage. In the direct experiment, both AlphaFold models and crystals from the PDB database were not modified. Standard docking procedures used programs such as AutoDock-4, ICM, rDock, and PLANTS, which differ in algorithms and scoring functions. The efficiency of docking in the model was assessed using ECR (exponential consensus ranking) and PRC (pose/ranking consensus) approaches.[113,114] Before starting the experiment, the authors compared the AlphaFold models and their crystallographic analogues using pLDDT and three different RMSD values. For some models, the binding site was blocked by other parts of the protein, hindering correct docking. Nuclear receptors, for example, ESR1, ANDR, and PRGR, can exist in various conformational states, but in AlphaFold models, predominantly only one conformation corresponding to the agonist-bound state is realized. For such cases, correct PDB database analogues for

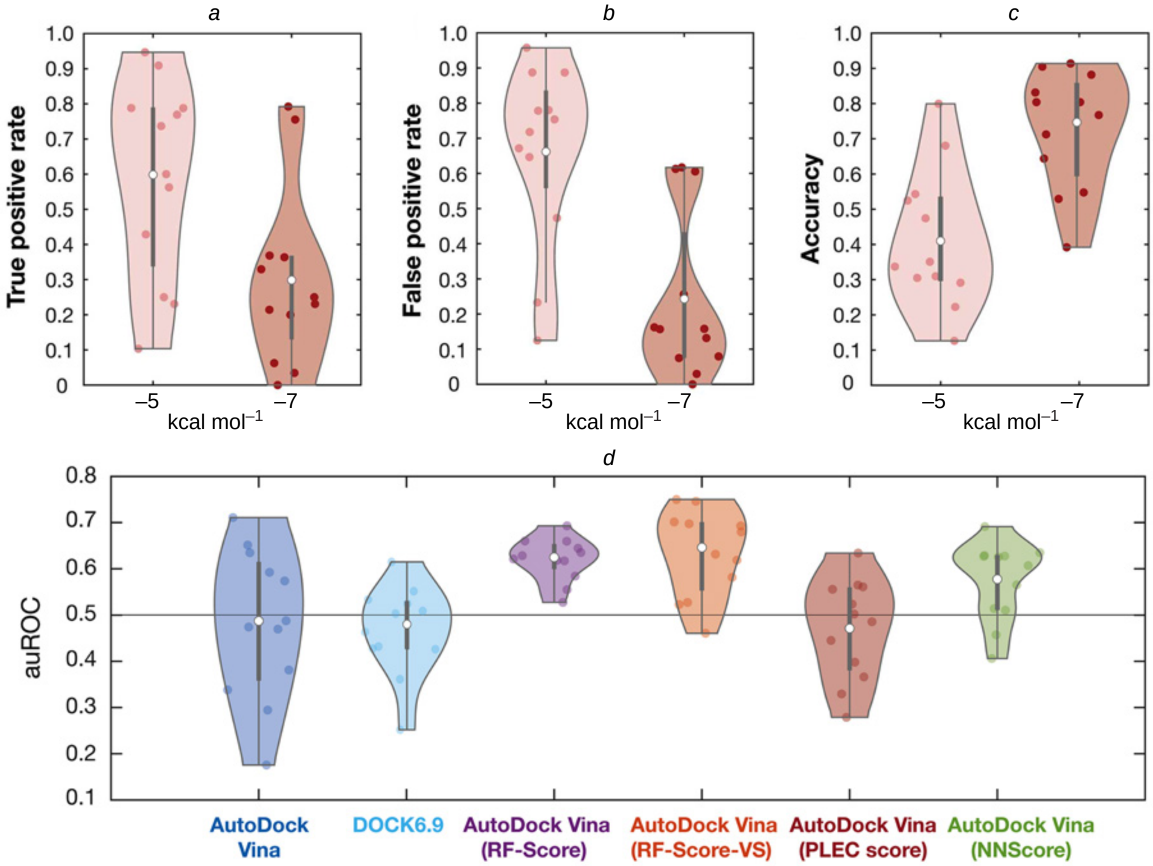


Figure 4. Distribution of various categories of tested molecules depending on the threshold value of the scoring function (*a–c*); modelling results (represented by auROC values) obtained using different scoring functions (*d*). The figure is published under CC BY 4.0 license.

comparison were chosen. In the AlphaFold docking, the ICM program showed the best results, but overall, the prediction quality was low [EF1-ECR = 8.8 and EF-PRC = 8.9, HR_{avg} = 0.16; EF1-ECR is average enrichment factor at 1% (EF1) by exponential consensus ranking; EF-PRC is average enrichment factor by pose/ranking consensus; HR is hit rate]. Many targets had EF < 3. For particular examples, the EF value was close to zero, with an average RMSD for ligand positions compared to docking results being 4.64 (in the case of cyclooxygenase-1, the RMSD was >10).

The docking results for ligands with AlphaFold models compared to crystals from the PDB database are presented in Table 4. It is evident that AlphaFold models are inferior to actual crystals, and this trend is characteristic of all four programs used for modelling. Comparable results were obtained for targets PRGR, PTN1, DRD3, and KITH, while for UROK, KPCB, ANDR, FABP4, ADRB2, and PYRD, modelling using PDB data showed better classifying ability, reflected in higher EF1-ECR values. The RMSD value for template ligands in the case of actual crystals, unlike AlphaFold models, was significantly lower ($RMSD_{avg}$ = 1.25). As in other publications, it was noted that the quality of the peptide backbone positions for the selected proteins is generally comparable to experimental data. This is not the case for the positions of side chains, especially in the binding pocket area, which undoubtedly affects the quality of the prediction in the context of virtual docking.

Based on the conducted experiment, the authors [112] concluded that the original AlphaFold models are mostly unsuitable for typical docking procedures, and this correlates with the observations of other researchers,[98, 115] but preprocessing of such structures generally improves the prediction quality. In particular, the crucial role of water at the ligand-binding interface was demonstrated using HSP90 as an example: the modelled solvent molecule positions allowed for an almost 8-fold improvement in RMSD between the true position and the docking result (RMSD values were 0.8 and 6.3 Å, respectively, in the presence and absence of water).

Hekkelman *et al.*[116] described the AlphaFill method for the automatic preprocessing of AlphaFold models, specifically aimed at improving the quality of docking. The authors noted that AlphaFold models are not designed to predict the positions of molecules that do not structurally relate to the peptide backbone and amino acid side chains. For example, haemoglobin should be considered in a complex with heme, zinc fingers should be considered with zinc atoms; these cofactors or coenzymes provide many DNA-binding proteins with stability and functionally significant spatial organization. Metalloproteinases should be analyzed with metal atoms in the active site, which dictates their catalytic activity. Similar conclusions can be made for other classes of protein molecules, for example, kinases, in which the binding of ATP molecules determines the enzyme conformation. Such compounds and atoms were referred to as transplants. The proposed algorithm allows for the transfer of necessary structures into AlphaFold models in the cases where their direct analogues are found in the PDB database.

According to the statistical processing performed in the cited study,[116] among the most frequently encountered transplants, one can highlight nucleotides (ATP, adenosine diphosphate, adenosine monophosphate, guanosine diphosphate, guanosine

Table 4. Docking results for ligands with AlphaFold models (AF) compared to crystals from the PDB database.[112]

Target	EF1-ECR PDB	AF	EF-PRC		Differences in the binding site structures of the AF model and the corresponding crystal from the PDB database
			PDB	AF	
ABL1	25.3	16.0	26.4	19.5	Asp^{381} is directed inward towards the binding site. There are minor differences in the conformation of the glycine-rich loop
PNPH	37.1	18.6	34.9	17.9	Ser^{33} has differences in the position of the OH group, which is submerged 2.66 Å into the pocket
ADRB2	24.5	3.4	23.4	2.5	Minor changes in the side chains of Asn^{1293} and Ser^{1203}
IGF1R	18.3	7.5	38.6	10.1	The DFG motif is located at the exit of the binding site. Gly^{1125} is positioned 4 Å away in the AlphaFold model.
CDK2	12.8	10.2	16.3	10.9	The side chains of Lys^{89} and Phe^{80} are directed inward towards the pocket, limiting its volume
COX1	3.4	1.3	5.8	2.5	The side chain of Phe^{518} is directed inward towards the pocket
PRGR	9.2	12.6	17.3	18.3	Trp^{755} is mirrored. There are differences in the position of the Gln^{725} side chain: the OH group is distanced by 2.45 Å
ANDR	9.0	0.0	13.5	0.0	Differences in the positions of the side chains of Gln^{711} and Thr^{877}
LFA1	10.9	2.9	11.6	0.0	The α7 helix (Asp^{297} : ^{1306}) is directed inward towards the binding pocket, limiting its volume
PTN1	29.5	29.5	23.9	21.3	The side chains of Asp48 and Asp181 are turned inward towards the pocket
UROK	25.9	2.5	47.0	2.5	The side chains of Asn^{322}, Ser^{323}, and Thr^{324} are directed inward towards the pocket (average RMSD = 2.28 Å).
FABP4	22.1	0.0	26.4	0.0	The side chain of Phe57 is directed inward towards the pocket (average RMSD = 1.6 Å).
KPCB	45.3	11.8	53.8	1.9	The *C*-terminal residues Cys622 : His636 are significantly shifted towards the binding site, altering its topology. The chain of Phe353 is located at the exit from the pocket
HSP90	0.0	0.0	0.0	0.0	Significant differences in the three-dimensional structure of the segment Asn^{106} : Gly^{137} near the pocket. The important water molecules for ligand binding are absent
ESR1	34.3	8.3	29.7	10.2	Minor differences in the conformation of the side chains of Met^{421} and His^{524} (shift towards the pocket)
DRD3	3.2	10.4	5.0	8.5	The side chain of Ser^{192} slightly extends out of the pocket. The radical of Trp^{369} is mirrored
KITH	22.1	22.1	20.0	20.7	Minor differences in the side chains of Arg^{53} and Arg^{61}
PDE5A	17.0	10.3	23.2	14.4	The side chain of Tyr^{664} significantly exits the pocket, whereas in the experimental structure, it interacts with the amino acids of the binding site. The side chains of Gln^{817} and Met^{816} are inversely rotated
FA7	47.1	13.1	48.0	23.2	Differences in the positioning of Lys^{189}
HXK4	5.5	1.1	15.2	0	The side chains of Ser^{64} : Phe^{66} are oriented towards the interior of the pocket, noticeably narrowing the space available for ligand binding. The side chain of Tyr^{214} is also directed inward into the site
PYRD	27.7	3.6	25.5	3.34	Minor changes in the positioning of the Arg^{136} and Tyr^{147} side chains. The side chain of Leu^{68} is directed into the pocket, contrary to experimental data. The side chains of His^{56} and Thr^{360} are inverted

triphosphate, uridine-5′-diphosphate), cofactors [coenzyme-A, flavin adenine dinucleotide, flavin mononucleotide, glutathione, heme, nicotinamide adenine dinucleotide (NAD), pyridoxal phosphate, *etc*.], and metal ions (Ca^{2+}, K^{+}, Mg^{2+}, Na^{+}, Zn^{2+}).

In the first stage, a search was conducted for template three-dimensional protein structures meeting two criteria: homology of at least 25% and at least 85 aligned residues (using the BLAST algorithm in the LAHMA program). Information about cofactors was obtained from the CoFactor database.[117] A total of 2694 structures were transplanted, which accounted for 95% of all ligands (excluding xenobiotics) in the PDB database. Spatial alignment of complexes was performed using standard techniques according to the position of C$_{\alpha}$ atoms, with the quality of overlay being assessed by RMSD values. The protein atoms located within 6 Å of the cofactor position made up the transplantation area and were aligned separately. As a result, 586 137 structurally modified AlphaFold models were obtained, with the total number of transplants exceeding 12 million. The initial validation of the AlphaFill algorithm was performed using proteins with 100% homology and the LEV (local environment validation) comparison function, which compares the cumulative RMSD of cofactors and the nearest amino acids in the obtained models with experimental data (Fig. 5*a*). The results presented in Fig. 5*b* indicate a relatively high correlation between LEV values and local RMSD. Overall, as expected, with increasing homology among amino acids that make up the pocket, the local RMSD decreases, despite a quite significant confidence interval (Fig. 5*c*). Using scoring functions defined based on the overlap of van der Waals volumes of atoms for polyatomic molecules after transplantation (TCS is transplant clash score, Fig. 5*d*), it was shown that TCS values correlate with LEV function values (Fig. 5*e*).[116]

For the optimization of complex geometry with high overlap coefficients (Fig. 5*f*), the authors[116] used the YASARA local minimization algorithm.[118] It is noted that after minimization of complexes with original TCS values close to zero, there is a

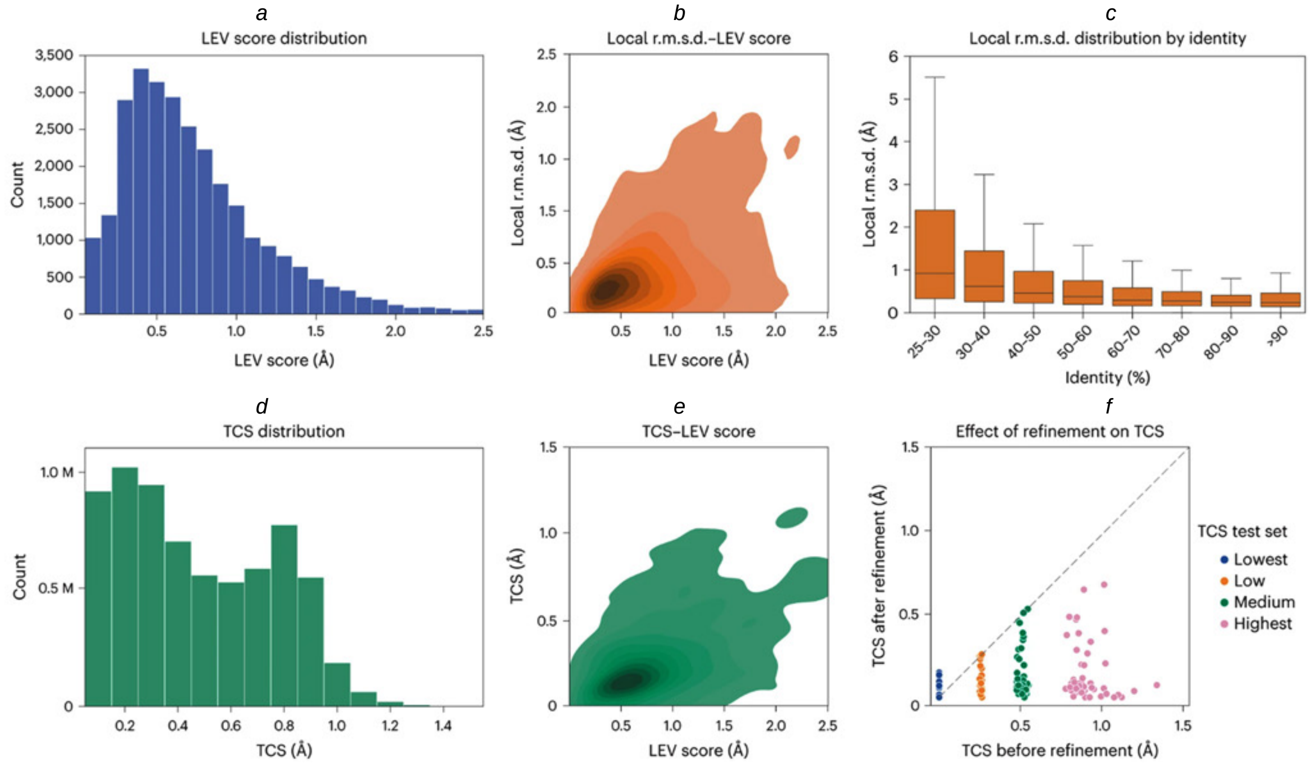


Figure 5. Distribution of LEV scores for validation examples (N = 28 619), data for 408 transplants with LEV > 2.5 are not shown (*a*); correlation between LEV values and local RMSD (*b*); dependence of binding site homology on local RMSD (*c*); assessment of volume overlap after transplantation (*d*); TCS – LEV dependence (*e*); comparison of TCS before and after local minimization for several sets of transplants (50 per group) with different initial values of this parameter (*f*).[116] The figure is published under the CC BY 4.0 license.

slight increase in the overlap, characteristic of cases where the cofactor does not have nearby protein atoms after transplantation. Under the action of the force field, the site 'compresses,' which, considering the intermolecular components of the field, leads to partial overlap of atom volumes. In other cases, the minimization procedure significantly lowered the TCS values. An example of optimization using the AlphaFill algorithm for myoglobin is shown in Fig. 6. Besides myoglobin, examples of transplants for zinc-dependent sites and kinase enzymes were discussed in the same paper.[116] Although the authors did not provide examples of algorithm validation in the context of docking of small molecules, it can be expected that in some cases, especially when the ligand affinity is, among other factors, determined by interaction with a cofactor, docking results will improve compared to the original model.

Liang *et al*.[119] thoroughly investigated the role of JMJD8 as an oncogene, which belongs to the JMJD protein family containing the Jumonji C (JmjC) domain in its structure. Proteins of this family can catalyze histone demethylation, similarly to HDAC (histone deacetylase-3) (deacetylation) and KDM (lysine-specific demethylase) (demethylation). However, such enzymatic activity was not demonstrated for JMJD8, likely due to mutations in the JmjC domain, while the N-terminal domain provides JMJD8 localization in the endoplasmic reticulum, indicating its role in folding. Besides, JMJD8 directly interacts with partner proteins, such as PKM2, leading to accelerated glycolysis. Specifically, it was found that the expression of this gene correlates with immunosuppression, DNA repair, and the activity of the CD276 protein, which is involved in regulating the T-lymphocyte immune response. Although the authors did not detect demethylase activity of JMJD8 in any of the conducted tests, they observed an association with several methyltransferases of other types.

Analysis of the gene expression profile using the cMap program,[120] particularly for the JMJD8 gene, in various tumour cells identified 26 molecules the impact of which induced

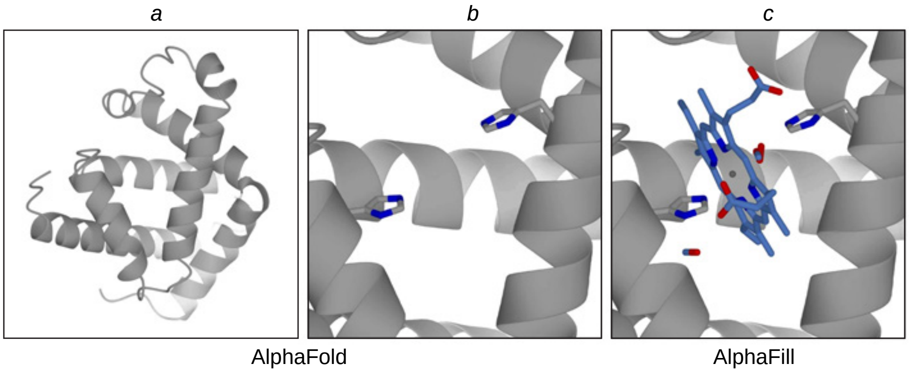


Figure 6. Original structure of the AlphaFold model (AF-P02144) (*a*); unoptimized positions of key histidine residues for heme structure positioning, with nitrogen atoms highlighted in blue (*b*); heme transplant in the AlphaFill model with CO_2 and O_2 molecules, carbon atoms highlighted in light blue, nitrogen atoms are in blue, and oxygen atoms are in red (*c*).[116] The figure is published under the CC BY 4.0 license.

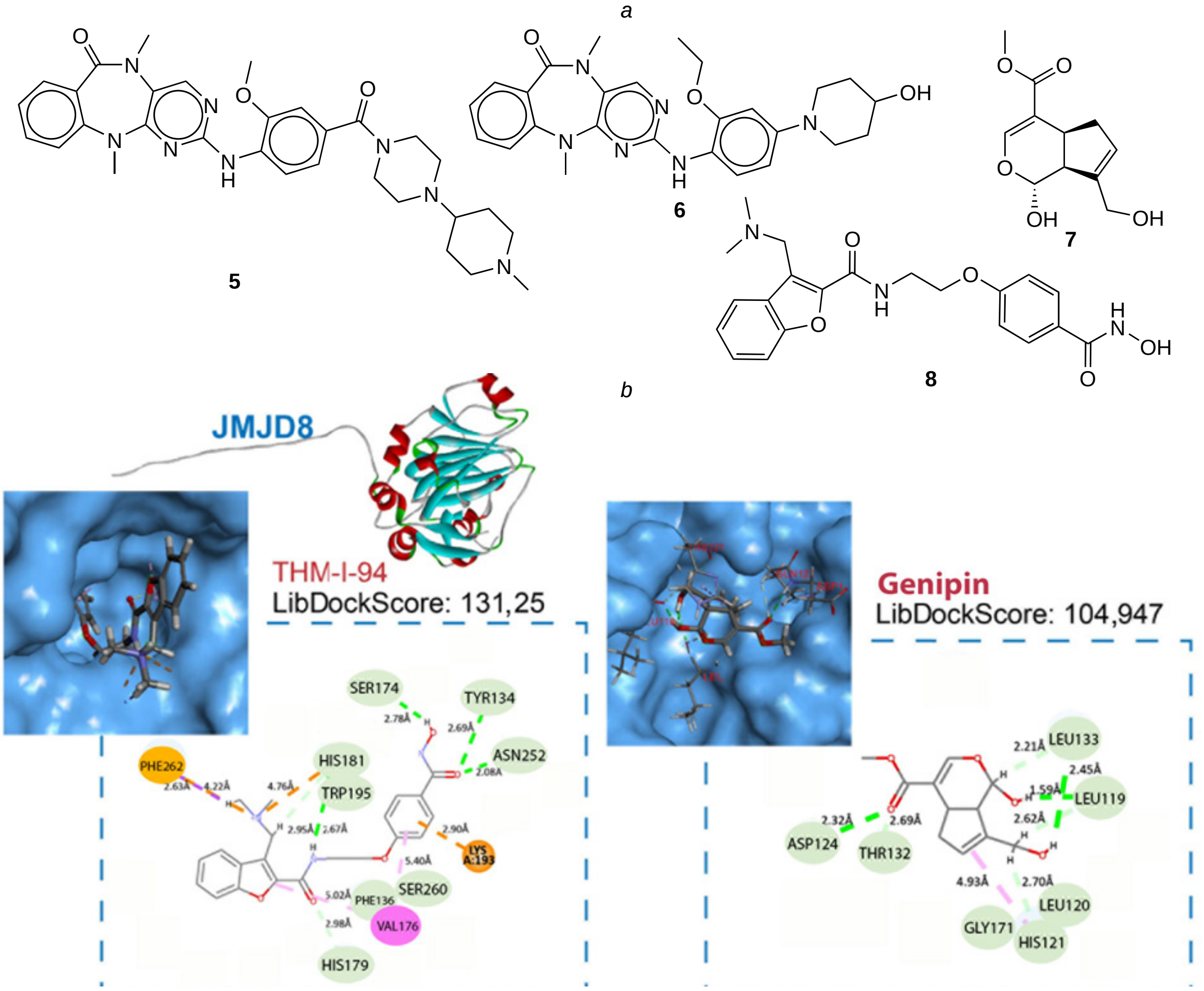


Figure 7. Structures of potential JMJD8 inhibitors (*a*) and docking results for genipin (**7**) and THM-I-94 (**8**) in the AlphaFold model (*b*).[119] The figure is published under the CC BY 4.0 license.

statistically significant changes in the expression of this gene. Among these, 6 molecules were categorized as HDAC inhibitors. In this context, Liang *et al.*[119] used the AlphaFold JMJD8 model for docking of four molecules [XMD-1150 (**5**), XMD-892 (**6**), genipin (**7**), and THM-I-94 (**8**), Fig. 7*a*] to predict their potential direct binding to JMJD8. The docking procedure, identification of the potential binding site, and preprocessing were carried out using the Discovery Studio v4.5 program (LibDock module).[k] The modelling results (Fig. 7*b*) showed that XMD-1150 and XMD-892 are incapable of interacting with the protein, while genipin and THM-I-94 demonstrated relatively high scoring functions (LibDockScore: 104.95 and 131.25, respectively).

Since the cited paper[119] does not provide detailed information about the specific features of the computer experiment and lacks experimental validation of the JMJD8-targeted mechanism of action for the selected molecules, assessing the role of the AlphaFold model in this work is quite challenging. It is worth noting that THM-I-94 contains a hydroxamic acid fragment typical of HDAC inhibitors, which interacts with the Zn^{2+} cation in the HDAC binding site, and the JMJD8 binding site visually resembles the catalytic site in the HDAC structure. This indirectly suggests the possibility that the molecule interacts with the studied protein. However, since, as mentioned earlier, the original AlphaFold models do not have cofactors in their structure, the docking results can be questioned if the authors did not perform a proper preprocessing procedure, which is not not mentioned in the cited publication.[119]

To discover new selective inhibitors of the enzyme OfHex1 (*O. furnacalis*), which participates in the hydrolysis of terminal *N*-acetyl-D-hexosamine residues in *N*-acetyl-β-D-hexosaminides, as insecticides, Satti *et al.*[121] described an approach based on the AlphaFold model. This work is related to agrochemistry, but the approaches to studying the mechanism of action and initial optimization of the structure of primary hit molecules are similar to those used in medicinal chemistry. In the first stage, to obtain comparative characteristics, the authors used available information for the specified enzyme from three organisms — *O. furnacalis*, *Homo sapiens*, and *T. pretiosum*. For enzymes from the first two organisms, the required crystallographic data were found in the PDB database: 3NSN (2.10 Å) and 1NP0 (2.5 Å). According to analysis, the overall amino acid homology was 40.11%, and the all-atom RMSD = 1.21 Å. In the case of *T. pretiosum*, an AlphaFold model was used (pLDDT = 86.75), with the Molprobity score[122] being 1.61 in the 92nd percentile ($N = 27675$, 0–99 Å) and the volume overlap coefficient being 1.75 in the 99th percentile ($N = 1784$, all resolutions). The homology with the nearest experimental complex (5Y1B) is characterized by corresponding values of 36.41% and 0.75 Å. Then, crystal preparation was

[k] https://www.computabio.com/discovery-studio-libdock-tutorial.html (access 28.03.2024).

carried out using the Maestro program, in particular, water molecules more than 5 Å away from the protein and ligand atoms were excluded from consideration, hydrogen atoms were added, and missing chain sections were built in the PRIME module. Complex minimization was performed using the OPLS4 force field.

Subsequently, the authors[121] compiled a library of commercially available molecules containing moieties typical of insecticides. In total, more than 20 000 molecules were selected, which underwent standard preprocessing using the LigPrep module. For each structure, no more than 32 starting three-dimensional representations were generated for docking (totaling 44 943 conformations). To validate the docking models, the authors predicted poses for ligands from the aforementioned crystals [TMG-chitotriomycin (**9**) and NAG-thiazoline (**10**) for 3NSN and 1NP0, respectively]. It was shown that using the constructed models, the true poses of the mentioned ligands are reproduced (Glide score = −12.95 and −6.53 kcal mol^{-1} for *O. furnacalis* and *Homo sapiens*, respectively.) In the AlphaFold model for a berberine derivative (**11**), a Glide score of –6.12 kcal mol^{-1} was obtained (*T. pretiosum*) (Fig. 8 *a*). The predicted binding mechanisms of the mentioned ligands are presented in Fig. 8 *b*. For docking of experimental structures of potential ligands, the Glide module was used (XP mode, that is, extra precision), and for the three poses with the best Glide scores, the free binding energy was evaluated using the MM-GBSA method (molecular mechanics with generalized Born and surface area).[123] The MD method in the Desmond module (Schrödinger LLC) was then applied to refine the binding mechanism and optimize the potential energy (using the OPLS4 force field), with the TIP3P scheme being used for water molecules, and system neutralization was achieved using Na$^+$ ions. Details of this experiment can be found in the original publication.[121] Ultimately, for each conformation, the 5 most probable MD results were selected. According to the modelling data (see Fig. 8 *a*), the active conformations for molecules **9** and **10** correspond to experimental data, while for compound **11** significant differences are observed, despite a high degree of spatial homology for the pockets.

As a result of docking of selected structures of potential ligands, 15 compounds (18 conformers) were identified, which are predicted to bind better than the control molecule **9** exclusively to OfHex1 (Glide score ⩽ –12.95 kcal mol^{-1}). Examples of molecules with the best Glide scores (**12**–**15**) are presented in Fig. 9.

The optimal poses for all 18 conformers were analyzed using the MD method in two stages (5 and 40 ns). After the first stage, 5 molecules were selected for which the RMSD over the trajectory did not exceed 3 Å. After the second stage, the greatest stability of complexes was predicted for three molecules — **9**, **13**, and **15**. The authors[121] did not provide the results of biological testing for the selected compounds, which does not allow for an assessment of the efficiency of their approach. Furthermore, no attention was given to the AlphaFold model, and possible reasons for the discrepancy in active conformations in the case of *T. pretiosum* were not explained.

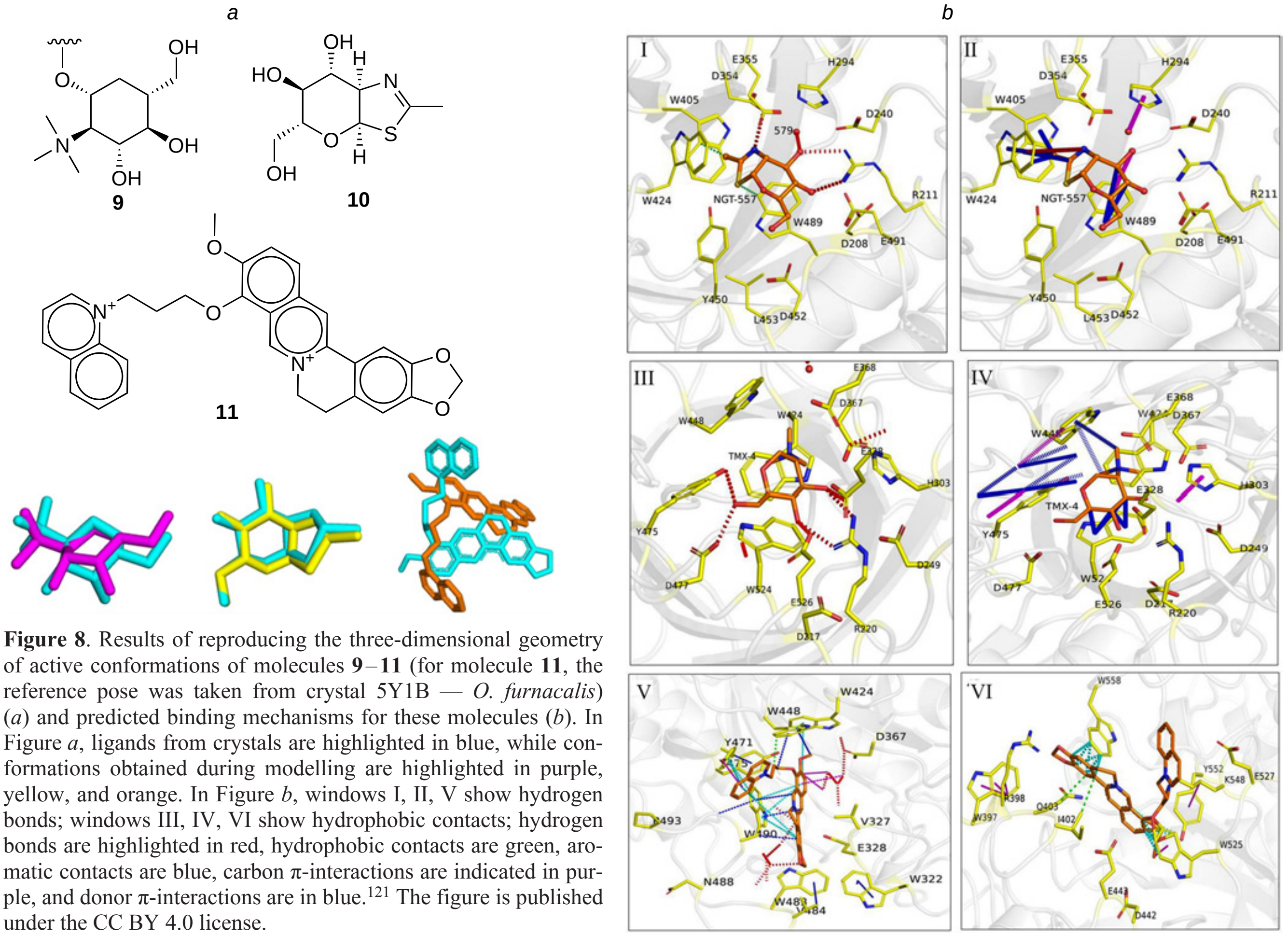
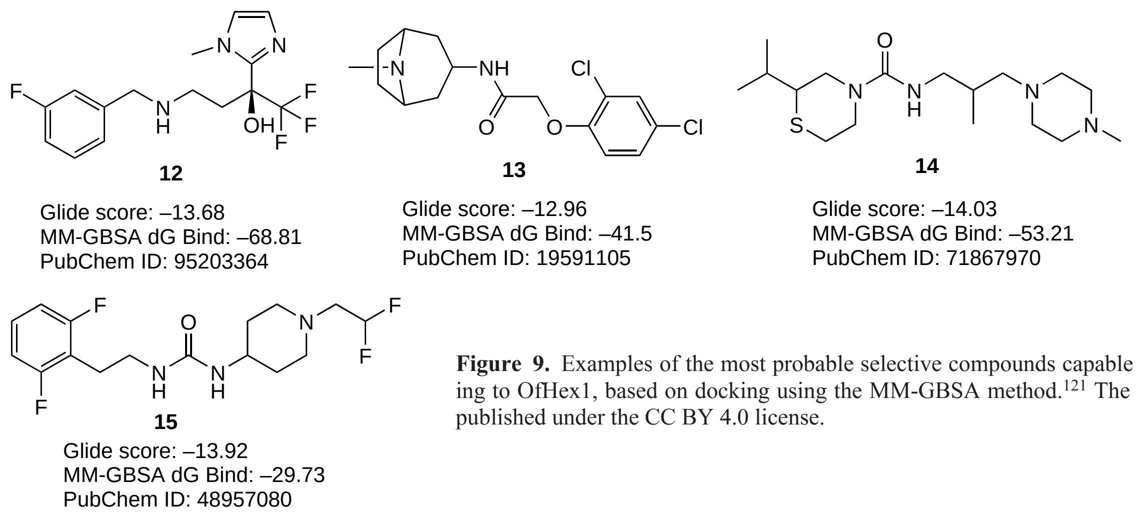


Figure 8. Results of reproducing the three-dimensional geometry of active conformations of molecules **9**–**11** (for molecule **11**, the reference pose was taken from crystal 5Y1B — *O. furnacalis*) (*a*) and predicted binding mechanisms for these molecules (*b*). In Figure *a*, ligands from crystals are highlighted in blue, while conformations obtained during modelling are highlighted in purple, yellow, and orange. In Figure *b*, windows I, II, V show hydrogen bonds; windows III, IV, VI show hydrophobic contacts; hydrogen bonds are highlighted in red, hydrophobic contacts are green, aromatic contacts are blue, carbon π-interactions are indicated in purple, and donor π-interactions are in blue.[121] The figure is published under the CC BY 4.0 license.



12

Glide score: –13.68
MM-GBSA dG Bind: –68.81
PubChem ID: 95203364

13

Glide score: –12.96
MM-GBSA dG Bind: –41.5
PubChem ID: 19591105

14

Glide score: –14.03
MM-GBSA dG Bind: –53.21
PubChem ID: 71867970

15

Glide score: –13.92
MM-GBSA dG Bind: –29.73
PubChem ID: 48957080

Figure 9. Examples of the most probable selective compounds capable of binding to OfHex1, based on docking using the MM-GBSA method.[121] The figure is published under the CC BY 4.0 license.

Lokhande *et al.*[124] used several AlphaFold models of viral proteins such as thymidylate kinase (A48R), DNA ligase (A50R), scaffolding protein D13 (D13L), palmitoylated EEV membrane protein (F13L), and bovine cysteine protease (I7L), to identify potentially active molecules against monkeypox virus (MPXV), which belongs to the Orthopoxvirus genus of the Poxviridae family. Recent studies have identified a significant homology (96.3%) between MPXV and the smallpox virus.[125] Therefore, the authors[124] selected the mentioned targets, for each of which, except A48R, known ligands exist: tecovirimat (ST-246) for F13L, TTP-6171 for I7L, rifampicin for D13L, and mitoxantrone for A50R. Preprocessing of the selected AlphaFold models was carried out using the Maestro program, following the standard protocol in the Protein preparation module. The SAVES (v6.0) program was used to construct Ramachandran plots, and the verification and analysis of three-dimensional structures were conducted in the ERRAT[1] and ProSA programs.[126] The molecules for docking were selected from the ChemDiv collection, PubChem, and DrugBank databases. A total of more than 206 000 structures were selected, which, after filtering procedures [REOS (rapid elimination of swill) and PAINS (pan-assay interference compounds) filters], were reduced to 171 000. Preprocessing and generation of starting three-dimensional representations for the final set of structures were performed in the LigPrep module (standard settings, OPLS4 force field).

As a result, more than 349 000 three-dimensional structures were obtained, the affinity of which was assessed based on the docking results. The binding sites for reference molecules were determined by analysis described by Lam *et al.*[127] and the results from the SiteMap module in the Schrödinger program (Fig. 10). To optimize the time and computational costs, the first stage of modelling was conducted in the high-throughput virtual screening mode and was followed by the application of the SP (standard precision) docking protocol for the top 10% of most promising compounds. In the final stage, for 10% of the most favourable conformations identified during the second stage, affinity was predicted in the XP mode. The likelihood of binding was assessed based on the Glide score. The stability of the predicted complexes and the affinity of ligands were analyzed using the MD method in the Desmond module, similar to what was done in the study cited above[121] (force field, OPLS 2005, cubic lattice 10 Å, 100 ns, TIP3P protocol, Nose-Hoover thermostat model, Martyna–Tobias–Klein barostat with isotropic coupling). Structural changes in the system were monitored based on RMSD and root-mean-square fluctuations of this parameter. Principal component analysis was used to visualize dynamic changes in the system. Complex energies were evaluated based on MM-GBSA calculations; the potential for interaction of molecules (virtual hits) with the binding site was assessed using calculations of the energies of the highest occupied and lowest unoccupied molecular orbitals, conducted using B3LYP and LDF functions[128] in the DMol3 module of the Discovery Studio program. The contribution of dispersion forces was calculated using DFT-D (dispersion-corrected density functional theory). AlphaFold models and corresponding pockets are presented in Fig. 10. In particular, the authors[124] noted that the results published by Lam's research group,[127] correspond to positions predicted in the SiteMap module. The main characteristics of the sites are presented in Table 5.

Based on the constructed Ramachandran plots, it was established that no more than 0.6% of amino acid residues (approximately 3 amino acids) in each model do not correspond to experimental data (except for I7L), with ERRAT scores [ERRAT is the cost function for non-bonded interactions between atoms (C, N, and O) in proteins] being 97.95, 91.47, 85.31, 95.04, and 87.50% for A48R, A50R, D13L, F13L, and I7L, respectively. Docking results showed that reference molecules have higher Glide scores than the promising molecules (virtual hits) from the test sample (Fig. 11). It is seen that for control molecules, Glide scores range from –4.92 to –4.52 kcal mol^{-1}, while for test molecules the maximum value of this parameter was –7.83 kcal mol^{-1} for molecule (I7L), and the minimum value was –11.27 kcal mol^{-1} (I7L, molecule **36**). MD results showed that the predicted complexes for reference molecules (except for the A48R target, which used an apo form) and compounds **16**–**40** are relatively stable. The principal component analysis results for the obtained trajectories provided the conclusion[124] that the binding of the most promising ligands causes significant changes in the structure of the apo form of the studied proteins, which presumably prevents the formation of a stable active conformation. For all investigated potential ligands, the energy difference between the highest occupied and lowest unoccupied molecular orbitals ranged from 0.0385505 to 0.2725508 eV, which, according to the authors, is related to the ability of molecules to interact with the pocket. Similar results were obtained using the DFT-D method. Like in several other

[1] https://www.doe-mbi.ucla.edu/errat/ (access 28.03.2024).

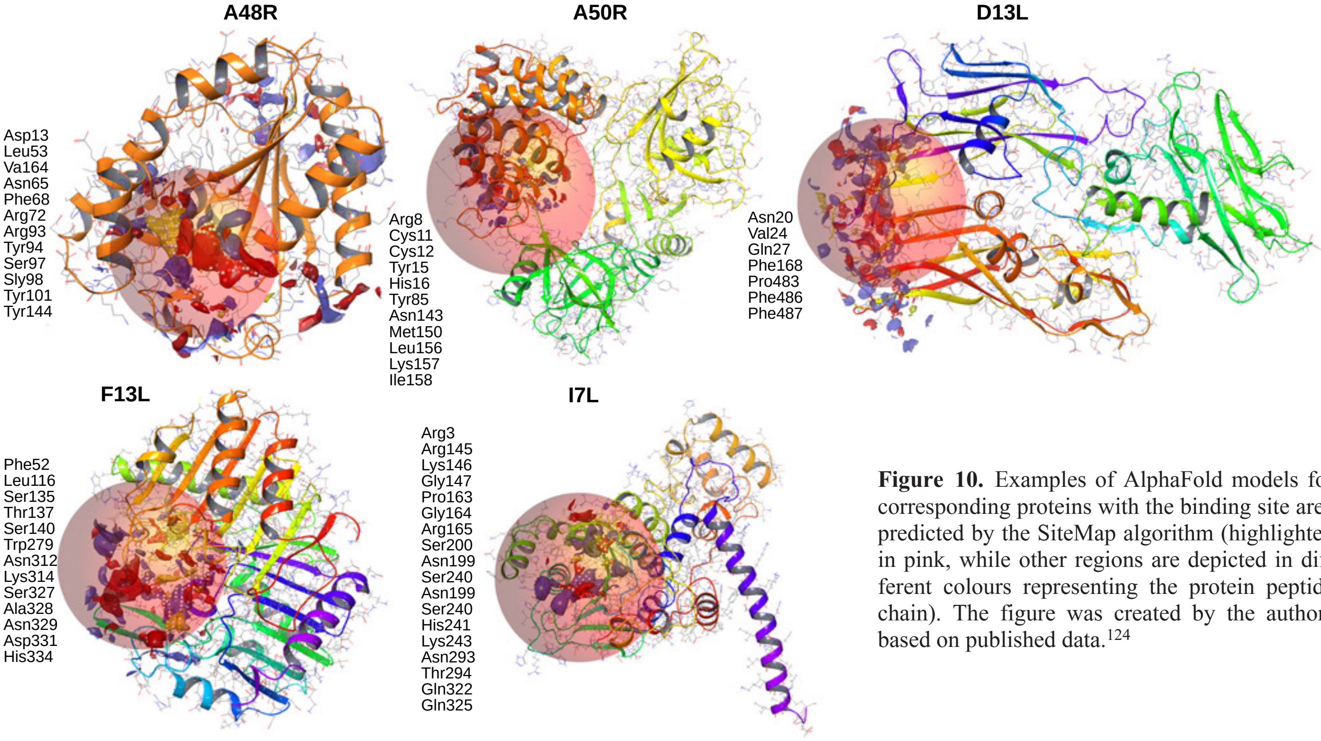


Figure 10. Examples of AlphaFold models for corresponding proteins with the binding site area predicted by the SiteMap algorithm (highlighted in pink, while other regions are depicted in different colours representing the protein peptide chain). The figure was created by the authors based on published data.[124]

Table 5. Key parameters of predicted pockets.[124]

Target	Druggable score†	Site score†	Site volume, Å^3	The numbers of amino acid residues forming the binding site
A48R	0.99	1.03	396.51	13, 14, 15, 16, 17, 18, 19, 37, 38, 39, 41, 53, 61, 62, 64, 65, 68, 72, 92, 93, 97, 98, 101, 102, 105, 107, 128, 129, 133, 134, 137, 142, 144, 145, 173, 175, 176, 177, 180
A50R	1.07	1.02	181.45	5, 8, 9, 11, 12, 15, 85, 143, 146, 147, 150, 156, 157, 158, 159
D13L	1.06	1.03	1134.99	1, 2, 3, 5, 6, 9, 10, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 28, 29, 30, 31, 32, 33, 35, 37, 39, 40, 41, 42, 55, 56, 57, 58, 59, 60, 62, 63, 64, 65, 161, 162, 164, 165, 167, 168, 169, 170, 171, 172, 217, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 228, 229, 230, 253, 480, 482, 483, 484, 486, 487, 488, 489, 491, 532, 537
F13L	1.04	1.01	444.53	52, 53, 55, 58, 86, 89, 112, 113, 114, 118, 120, 133, 135, 137, 139, 140, 144, 239, 246, 247, 248, 249, 279, 281, 282, 283, 312, 314, 327, 329, 331, 333, 334, 338
I7L	1.04	1.01	502.50	1, 3, 4, 8, 138, 140, 168, 238, 239, 240, 241, 243, 258, 260, 261, 262, 263, 264, 266, 275, 277, 278, 280, 281, 283, 284, 285, 286, 294, 295, 318, 319, 320, 321, 322, 323, 325, 326, 328

† Parameters obtained in the SiteMap software (Druggable score is the score for pocket properties and Site Score is the overall pocket assessment); Druggable score (DScore) and SiteScore functions are based on the same equation parameters, but have different coefficients. Druggable score better characterizes the ability to bind to drug molecules.

mentioned works, the cited study[124] does not provide biological testing results for the selected molecules.

Nussinov *et al.*[129,130] noted that the efficiency of using AlphaFold models depends on the task at hand, and their potential in fields such as the development of new drug molecules and medicinal chemistry is not fully clear for a number of reasons. Proteins are not static objects, especially near active sites, as is the case for many enzymes. Solvent molecules play an important role in the interaction of ligands with binding sites. Even single substitutions among the amino acids lining the pocket can have a significant impact on the affinity of small molecules. The authors emphasize that many enzymes exist in various states, citing kinases and 5-HT$_{5A}$ serotonin receptors as examples.

The reasons listed above, among other factors, do not allow AlphaFold structures to be considered as suitable models for conducting standard computer modelling procedures without additional modification. Researchers compare AlphaFold models to photographs, not to the living systems that proteins are complex dynamic entities whose degrees of mobility are far from being limited to two states (active and inactive). Identifying allosteric binding sites using AlphaFold models is likely the most challenging task, as allosteric sites may form only in certain states of the protein, for example, during interaction with a partner molecule, or depend on surrounding conditions, particularly pH. An example is the development of the allosteric inhibitor asciminib against the BCR-ABL1 oncoprotein for chronic myeloid leukemia.[131] However, ligand binding to an allosteric pocket does not always guarantee a clear effect due to poorly predictable conformational transitions. For instance, inhibitors of integrins aIIbb3 and a4b1 stabilize the active conformation, demonstrating properties of partial agonists, which posed a serious problem during clinical trials.[132]

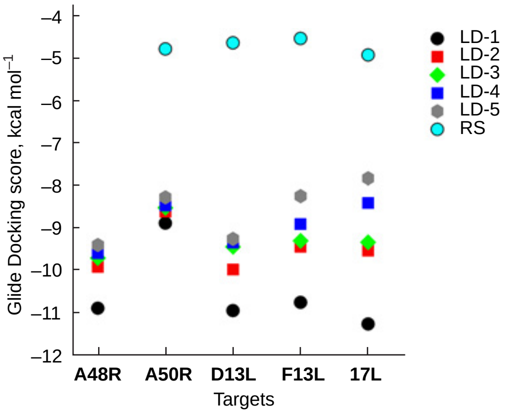


Figure 11. Docking results of reference structures (RS) compared to virtual hit molecules: LD-1 (**16, 21, 26, 31, 36**), LD-2 (**17**, **22**, **27**, **32**, **37**), LD-3 (**18**, **23**, **28**, **33**, **38**), LD-4 (**19**, **24**, **29**, **34**, **39**), and LD-5 (**20**, **25**, **30**, **35**, **40**). The figure was created by the authors based on published data.[124]

Considering the above, as one of the possible solutions to the listed problems, Nussinov *et al.*[129,130] suggested integrating information about functionally significant states of the protein into AlphaFold models, for example, about the kinase DFG-in/DFG-out conformation, using pharmacophores for the inactive, typically more stable state of the protein, which in most cases is realized in AlphaFold models. The authors specifically mentioned a publication[133] discussing an approach that takes account of various states of GPCR proteins in the AlphaFold algorithm.

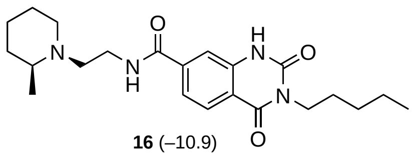
4.2. Design of novel small molecules

Recently, Ren *et al.*[134] published the results of a study in which they managed to identify the protein CDK20 as a new potential target for the drug therapy of hepatocellular carcinoma using the PandaOmics program.[135] In the absence of experimental data on the structure of the selected target, the authors applied the AlphaFold algorithm to construct its three-dimensional model. Using the generative platform Chemistry42,[136] a series of new

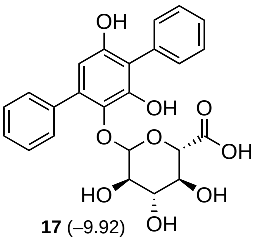
Structures 16–30

(the values in parentheses are scoring functions, kcal mol^{-1})

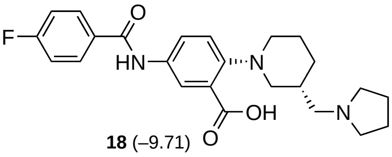
A48R



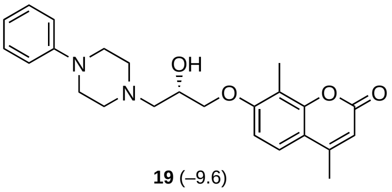
16 (–10.9)



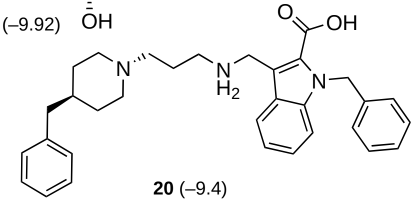
17 (–9.92)



18 (–9.71)

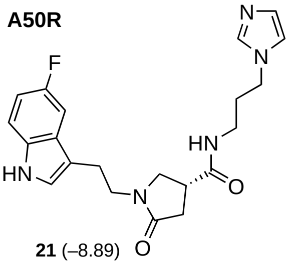


19 (–9.6)

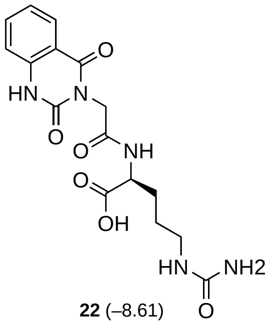


20 (–9.4)

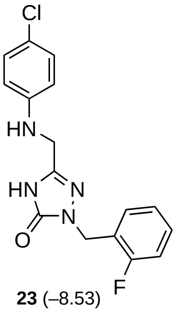
A50R



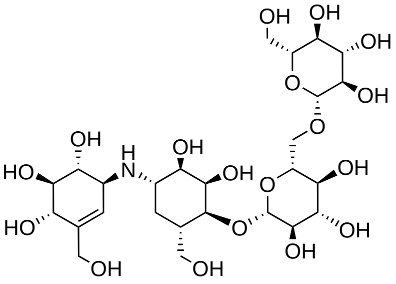
21 (–8.89)



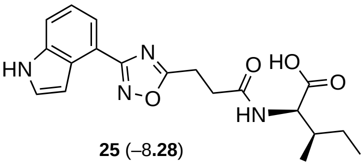
22 (–8.61)



23 (–8.53)

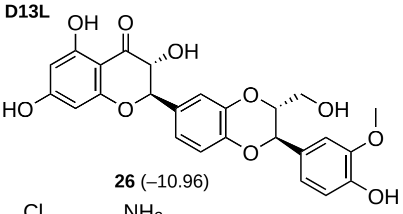


24 (–**8**.47)

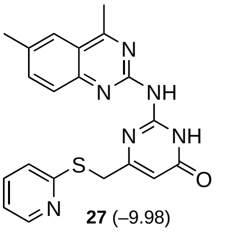


25 (–8.**28**)

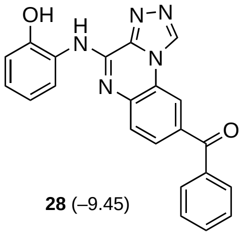
D13L



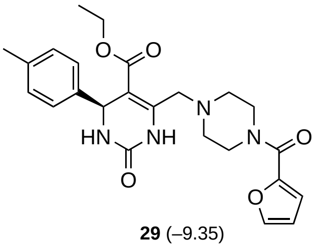
26 (–10.96)



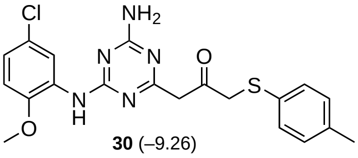
27 (–9.98)



28 (–9.45)

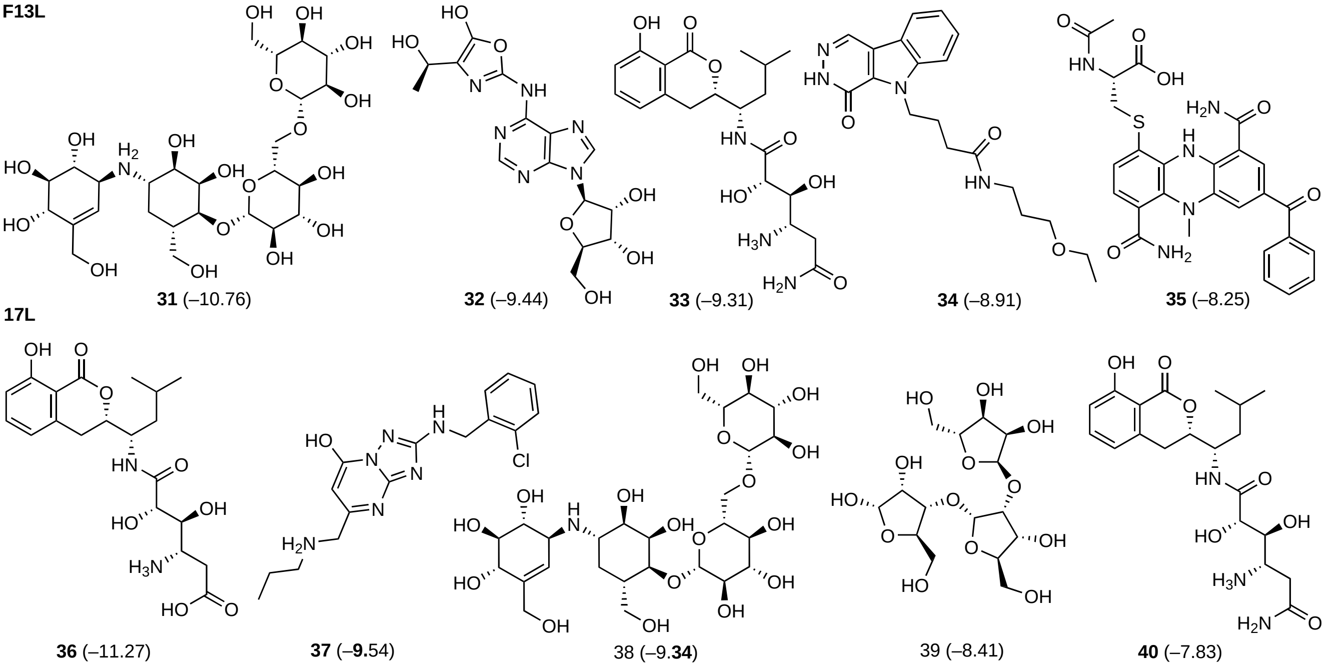


29 (–9.35)



30 (–9.26)

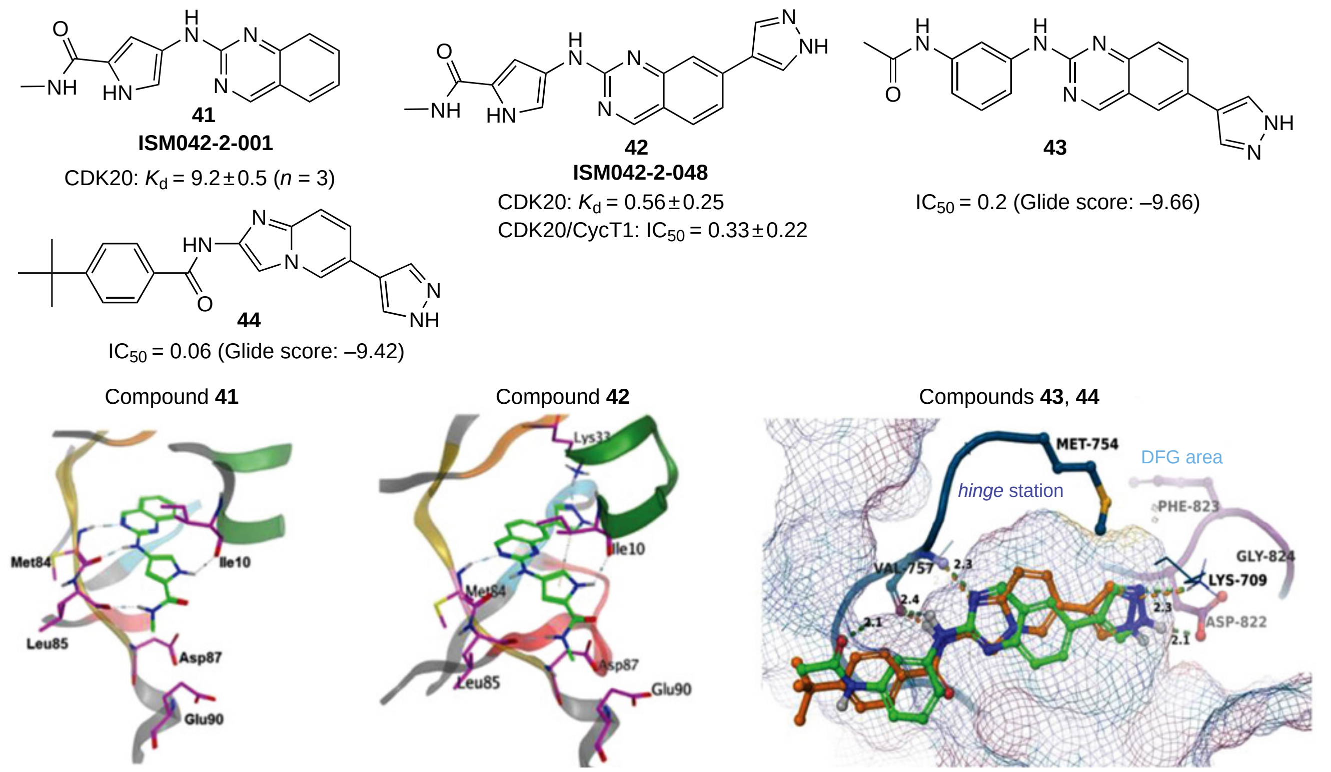
Structures 31 – 40
(the values in parentheses are scoring functions, kcal mol^{-1})



31 (–10.76) **32** (–9.44) **33** (–9.31) **34** (–8.91) **35** (–8.25)

36 (–11.27) **37** (–**9.**54) 38 (–9.**34**) 39 (–8.41) **40** (–7.83)

structures of potential ligands was proposed, some of which were synthesized and tested *in vitro*. While selecting molecules for synthesis, the researchers also relied on the results of docking (Fig. 12). The most active compound was **41** (ISM042-2-001, dissociation constant $K_d = 9.2\pm0.5$ μM, $IC_{50} > 6$ μM). The first experiment lasted 30 days; during this time, only 7 molecules were synthesized, six of which showed no activity. In the second stage of generation and synthesis, compound **42** (ISM042-2-048) was obtained, showing better results in biological testing ($K_d = 0.57\pm0.26$ μM, $IC_{50} = 33.4\pm22.6$ nM). In the second



41
ISM042-2-001
CDK20: K_d = 9.2±0.5 (n = 3)

42
ISM042-2-048
CDK20: K_d = 0.56±0.25
CDK20/CycT1: IC_{50} = 0.33±0.22

43
IC_{50} = 0.2 (Glide score: –9.66)

44
IC_{50} = 0.06 (Glide score: –9.42)

Figure 12. Examples of the structure of CDK20 inhibitors (**41**, **42**) and their closest known analogues: ASK1 kinase inhibitors (**43**, **44**), and docking results (IC_{50} and K_d values are given in μM, Glide score values are in kcal mol^{-1}); carbon atoms of the ligands are highlighted in green and orange, nitrogen atoms are in blue, oxygen atoms are in red, and hydrogen atoms are in white.[134] The figure is published under CC BY 3.0 license.

stage, only 6 molecules were synthesized, two of which were active. Apparently, the introduction of a pyrazole moiety into the molecule significantly influenced the activity, which is not surprising considering the numerous examples in the development of kinase inhibitors, where such acceptor moieties interacted with the lysine (in this case, Lys88) amino group either directly or through bridging water molecules. In the test using the Huh7 cell line expressing CDK20, compound **42** demonstrated relatively high activity ($IC_{50} = 208.7 \pm 3.3$ nM), while the activity against HEK293 was 1706.7 ± 670.0 nM (IC_{50}, selectivity index SI = 8.2). The authors noted that the constructed three-dimensional model (AF-Q8IZL9-F1-model_v1) did not allow for a correct docking procedure and was manually modified, for example, the *C*-terminal section ($Pro^{303}-Gly^{346}$) was removed. For modelling, from one to 302 amino acids were used, while the model corresponded to the DFG-in conformation, and the charged amino acids Asp^{87} and Glu^{90} were located in an area accessible to the solvent.

Ren *et al.*[134] were the first to demonstrate the effectiveness of using AlphaFold to create new active ligands against a kinase not previously described as a potential target for the therapy of hepatocellular carcinoma. It is specified that the new target is considered to be a protein that is 'targetable' (without explaining the term), has the status of 'new target' in the PandaOmics program, has not been addressed in any clinical trials in the last three years, and is not a target for known approved drug molecules. This definition, in our opinion, does not fully correspond to the concept of a 'new target.' The term 'targetable' is interpreted differently by medicinal chemists, cheminformaticians, and clinicians. For example, from the perspective of clinical pharmacology, 'targetability' means that the protein and the drug molecule acting on it have proven themselves during clinical trials as an effective strategy in treating a specific disease with acceptable side effects and a suitable clinical outcome, which contradicts the authors' definition. Moreover, the rationale behind the choice of a period during which no results of clinical research should be published (the last three years) is unclear. Likely, the authors[134] adhere to a definition appropriate in the field of medicinal chemistry, where 'targetable' means that examples of small molecules acting on the selected target are known, which *de facto* excludes the possibility of correctly applying the term 'new target.' Specifically, the cited paper[134] claims that molecule **42** contains a new fragment (based on the Tanimoto coefficient value) capable of binding to the hinge region, absent in known CDK20 inhibitors.

The metric mentioned for evaluating the novelty of structures is not typically used: novelty can only be assessed based on a thorough literature and patent search. The authors[134] have filed the patent application WO2023138412 (A1); however, close analogues of the presented structures were previously described as, for example, ASK1 kinase inhibitors (compounds **43** and **44**, see Fig. 12).[137] Essentially, in the structure of ISM042-2-048, there was an isosteric replacement of the phenyl group with a pyrrole ring while maintaining the inverted amide group while the pyrazole moiety, which in molecule **43** also contacts the lysine residue similarly to Lys^{88}, was moved to the 7-position of the benzopyrimidine. However, docking results for ISM042-2-048 show the pyrrole proton interacting with the carbonyl group of the peptide backbone of Ile^{10} through the formation of a hydrogen bond, absent in the case of the phenyl group. Given the above, it could be assumed that testing molecule **43** against CDK20 or compound **42** against ASK1 would provide a better evaluation of the uniqueness of the approach described by Ren *et al.*[134]

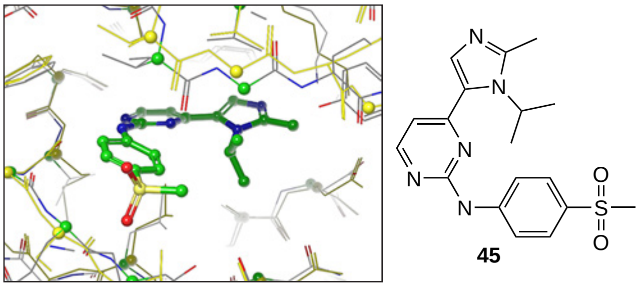


Figure 13. Spatial structure of the pockets of CDK20 (highlighted in yellow) and CDK2 (highlighted in grey) using molecule **45** as an example. Carbon atoms are marked in green, sulfur is in light yellow, oxygen is in red, and nitrogen is in blue. The figure was created by the authors based on published data.[134]

It should be noted that among the published crystallographic data, we discovered a co-crystal of the topological analogue **45** (AZD-5438, a non-selective inhibitor of CDK kinases)[138] in complex with CDK2 kinase (PDB: 6GUH, crystal resolution 1.50 Å). Alignment of 6GUH and AF-Q8IZL9-F1-model_v1 showed that the pockets of CDK20 and CDK2 have high homology (Fig. 13). In the case of AZD-5438, the imidazole acceptor nitrogen provides interaction with the lysine residue through a water molecule. Given that Ren *et al.*[134] did not present selectivity research results for ISM042-2-048, at least regarding the CDK kinase family, it can be speculated that comparable results could have been achieved using a simple homologous model. Despite the mentioned observations, in the studied research, the AlphaFold model allowed medicinal chemists, albeit not without expert modification of the modeled CDK20 structure, to develop hit molecules with inhibitory activity against CDK20 kinase.

Another study[139] reported by the same research group addresses the development of selective SIK2 kinase inhibitors using the AlphaFold model (selective SIK2 inhibitors are described in the literature, for example, ARN-3236, Ref. 140). Specifically, the authors compared their own homologous model, constructed based on the previously proposed binding mechanism of the non-selective SIK inhibitor MRIA9,[141] using the AlphaFold model, while details of the reproduction of the homologous model are not provided. It is noted that when overlaying the two models, significant differences (7.58 Å) are observed in the P-loop (Asn^{30}) in the binding sites (Fig. 14*a*). We reproduced the homologous model in accordance with the methodology described by Tesch *et al.*,[141] using the complex structure 7B30 as a template and the SWISS-MODEL program, and confirmed these differences.

Zhu *et al.*[139] performed docking simulations for four known ATP-competitive inhibitors with both the constructed homologous model and the model obtained from AlphaFold (AF-Q9H0K1-F1-model). It was indicated that the selected compounds interact with the hinge region; however, in the case of GLPG-3970 (**46**), the 3,4-dihydroisoquinolin-1(2*H*)-one moiety is located at the exit of the pocket (Fig. 14*b*), unlike the docking results obtained using the AlphaFold model, where this moiety is located near the gatekeeper area, and the methoxy group occupies a pocket formed by amino acids in this region (Fig. 14*c*).

We attempted to reproduce the reported results and conducted docking simulations of GLPG-3970 with both models. The constructed homologous model was not minimized and only underwent standard preprocessing, as was the case with the AlphaFold model. We precisely replicated the binding pose of

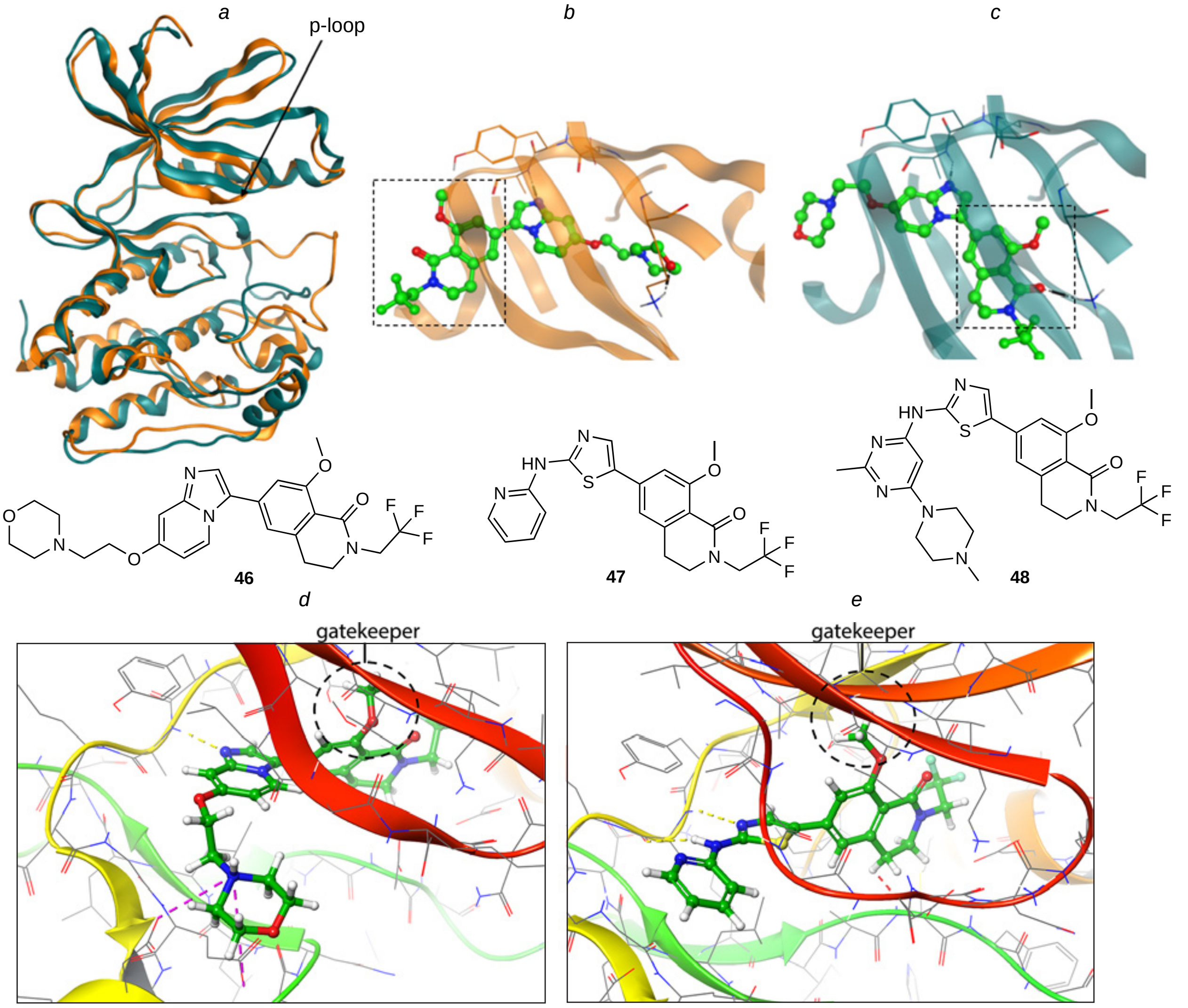


Figure 14. Visualization of the spatial overlay of the homologous model and the AlphaFold model (*a*), docking results of compound **46** with the homologous model (*b*) and with the AlphaFold model (*c*);[139] docking results of molecules **46** and **47** with the model constructed by the authors of this review (*d*, *e*). The methodology used for the construction is similar to that referenced in Ref. 139, which is also similar to that used by Tesh *et al.*[141] Dashed lines highlight ligand fragments interacting with the gatekeeper region; atom labelling is the same as in Fig. 13. The figure was created by the authors of the review based on published data.[139].

G-5555 in the SIK2 kinase binding site, but for GLPG-3970 in the case of the homologous model, we did not observe the differences indicated by Zhu *et al.*[139] Key contacts with the hinge region are maintained, and the oxygen of the carbonyl group interacts with the lysine residue (Lys^{49}). It is important to note that we obtained a similar pose for the hit molecule **46** described previously[139] (Fig. 14 *d*). Conversely, poses where the dihydroisoquinoline moiety is located at the exit of the pocket were obtained precisely during docking simulations with the AlphaFold structure. It should be noted that based on MD results, the authors considered the pose close to the one that was *a priori*ty in our computational experiment using the simple homologous model, but with a different torsion angle relative to the bond connecting the dihydroisoquinoline and thiazole moieties. The main focus in the previous publication[139] is the interaction of the methoxy group with the gatekeeper area, which, in particular, correlates with the results of structure–activity relationship analysis presented in that study. However, it remains unclear why the AlphaFold model better describes the patterns observed in the biological experiment. As a result, we were unable to confirm the significance of the AlphaFold model in this study. Nonetheless, using the Chemistry42 program, structures of new active compounds were generated, among which **47** and **48** demonstrated comparatively high activity *in vitro*: $IC_{50} = 23$ and 0.7 nM, respectively. In particular, high selectivity was noted for molecule **48** (SI = 24 and 200 for SIK2/SIK1 and SIK2/SIK3, respectively). At a concentration of 100 nM, no comparable activity was observed against other ten kinases, including AMPK kinase. The authors explained this by the presence of the Thr^{96} residue in the gatekeeper area of this kinase (unlike methionine in the SIK2 structure), affecting the binding of the methoxy group.

Mendes *et al.*[142] utilized the AlphaFold domain model of diacylglycerol kinase (DGK) and the molecular probe TH211 (**49**) (Fig. 15), which is specific to tyrosine and lysine residues in the DGK active site (RF001-sites),[143] to identify potential binding areas for small organic molecules across a range of ten DGK isoforms. To study this, chimeric DGK family proteins were created, incorporating the C1 domain (tandem C1A and

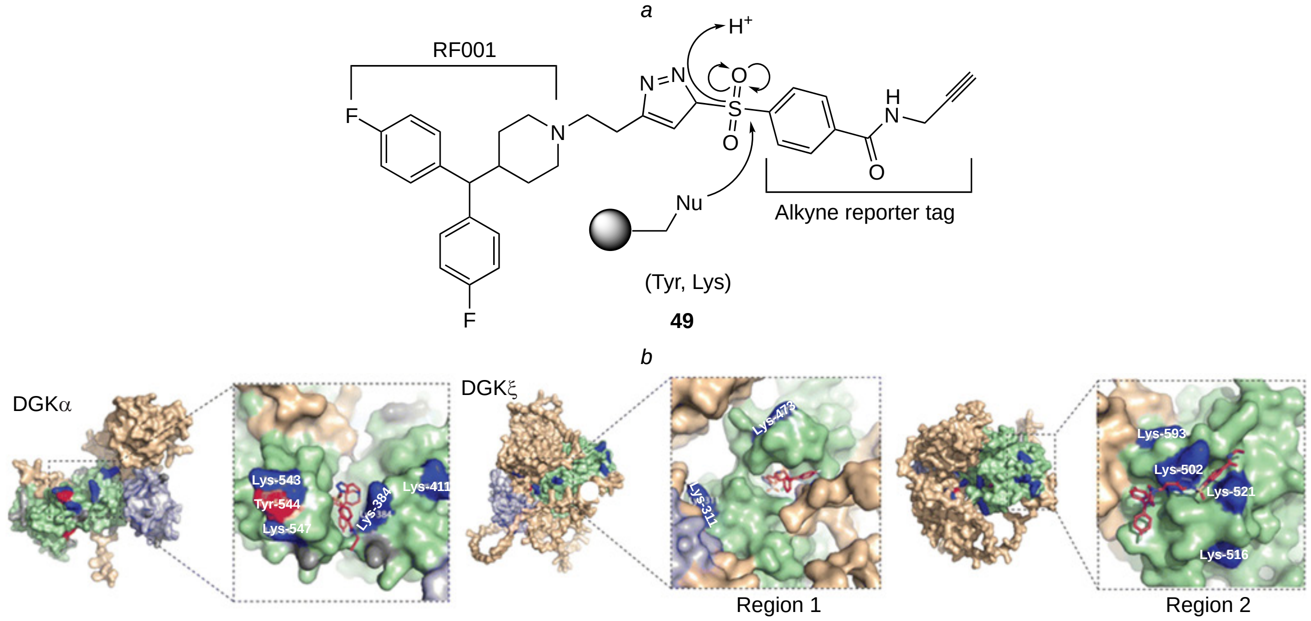


Figure 15. Structure of TH211 (*a*) and visualization of binding sites in AlphaFold models (*b*). Domains C1A and C1B are highlighted in light blue, while the catalytic domain (DGKα and DGKξ regions) is highlighted in light green. Modified amino acids Lys and Tyr are shown in dark blue and red, respectively; Lys and Tyr residues that are accurately predicted by AlphaFold and unmodified are marked in grey.[142] The figure is published under the CC BY 3.0 license.

C1B domains), which is responsible for regulatory protein–protein interactions with partner proteins and provides specificity towards the substrate (diacylglycerol). Biological experiments demonstrated that TH211 blocks the catalytic activity of all isoforms through covalent interaction, with binding sites being identified not only in the C1 domain. Using AlphaFold models, the authors then mapped pockets by tyrosine and lysine residues in all isoforms (for tyrosine and lysine residues, pLDDT > 70), including chimeric proteins. As a result, clusters containing modified amino acids were identified, for example, for DGKα (K384, K543, Y544, K547), DGKζ (K502, K516, K521, K593; K311, K473), DGKγ (K356, Y358, Y535, K542), and DGKδ (K271, K198, K337). Additionally, the close spatial arrangement of the C1 and catalytic domain in the DGKα structure predicted by AlphaFold correlates with experimental data. This approach provides structural information about potential small molecule binding sites to DGK family proteins and can be used for developing selective inhibitors. AlphaFold models have also been used by other authors for predicting binding sites, for instance, in the structures of phosphatase PPM1D/Wip1,[144] KRAS,[145] andsacsin.[146]

Li *et al.*,[147] using AlphaFold models for the protein PcOSBP (*P. capsici*), developed a series of new inhibitors–structural analogues of oxathiapiprolin (**50**–**53**) (Fig. 16), possessing fungicidal activity. As a control structure, the authors used crystallographic data (PDB: 1ZHY, 1.5–1.9 Å resolution) for the KES1 protein (*Saccharomyces cerevisiae*) from the aforementioned family with a low degree of homology (less than 30%) to PcOSBP in complex with ergosterol, cholesterol, and 7-, 20-, and 25-hydroxycholesterols. Sequence alignments were carried out using the MUSCLE Web program.[m] Initially, an AlphaFold model was built for the control protein, resulting in a comparatively low average RMSD (0.54 Å), based on which the authors concluded that the algorithm is suitable for modelling the spatial structure of PcOSBP. Despite somewhat contentious reasoning, using the MD method, the authors optimized the AlphaFold model in complex with oxathiapiprolin to achieve a stable conformation. For docking the synthesized compounds (both isomers), the LeDock program [148] was used. For each structure, 100 docking attempts were made, and the affinity of the ligands was assessed based on the LeDock score and visual analysis of the predicted interaction. The docking results are presented in Fig. 16.

It was shown that the binding mechanisms of the reference molecule and synthesized compounds coincide due to their high structural homology. Essentially, the authors [147] applied the scaffold hopping approach and modified the pyrazole moiety of oxathiapiprolin while preserving the key interaction with the lysine residue (Lys^{703}). The predicted binding energy values (ΔGPB) for *R*- and *S*-oxathiapiprolin were –24.73 and –24.59 kcal mol^{-1}, respectively (the authors attributed the different values, in part, to the presence of a hydrogen bond with the Asn^{767} residue in the case of the R-isomer). Comparable docking results were obtained for compound **53**: ΔGPB = –23.37 (*R*) and –22.32(*S*) kcal mol^{-1}. Indeed, the *R*-isomer of oxathiapiprolin showed higher activity against *P. capsici* than the *S*-isomer [149] [EC$_{50}$ = 0.17(R) and 0.66(S) µg L^{-1}; EC$_{50}$ is the half-maximal effective concentration], despite the relatively small difference in binding energies. The fungicidal activity of the synthesized molecules was investigated against *P. capsici*, *Peronophthoralitchii* (*P. litchii*), and *P. infestans*; at a concentration of 0.01 µg L^{-1}, compound **51** demonstrated a weaker inhibitory effect than the control molecule: 28.15, 32.62, and 44.15% respectively. Compound **52** showed stronger activity at the same concentration, namely 94.93, 84.60, and 95.9%, comparable to the activity of oxathiapiprolin. Compound **53** more selectively inhibited the growth of the mentioned organisms at a similar concentration (99.19, 70.67, 87.11%), which was particularly associated with the ethyl group, which, according to docking results, could occupy the hydrophobic section of the binding site.

[m] https://www.ebi.ac.uk/Tools/msa/muscle/ (access 28.03.2024).

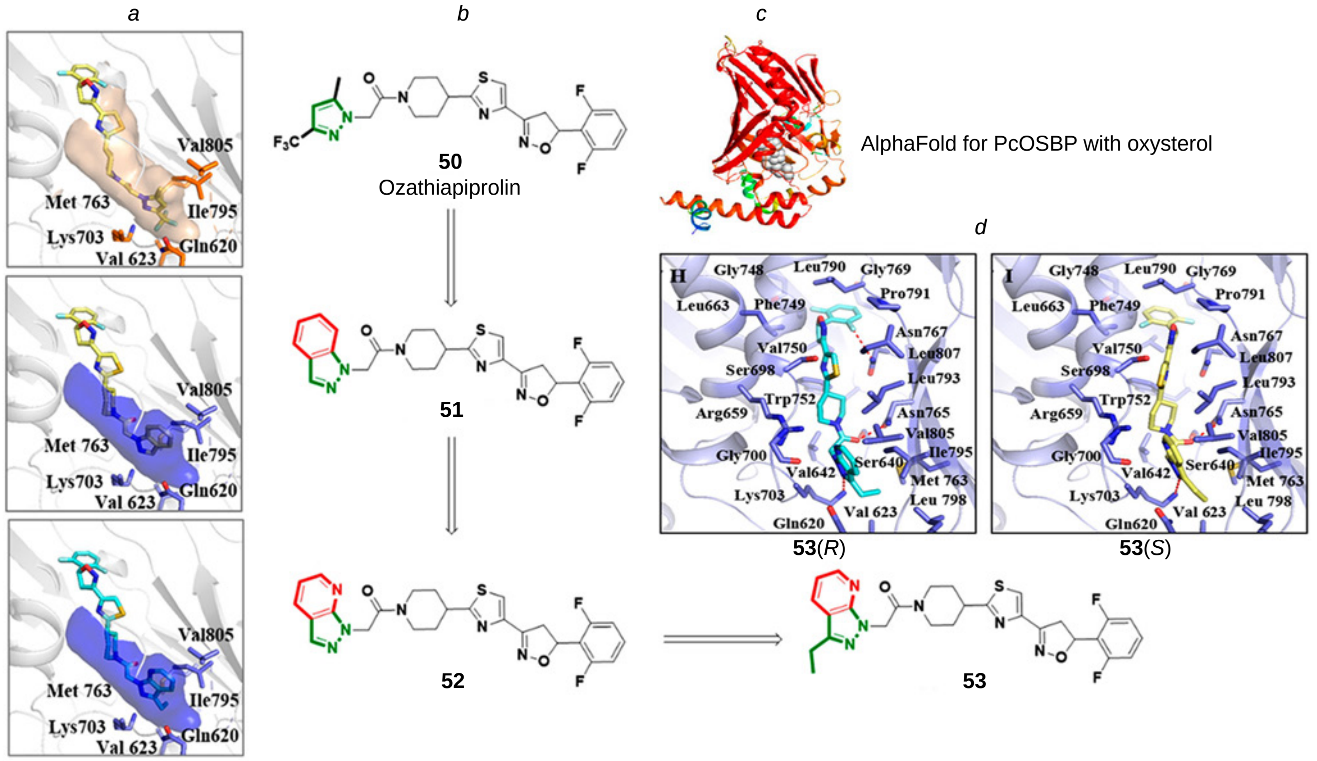


Figure 16. Docking results of compounds **50**–**52** (highlighted in yellow and blue) (*a*);[147] structures of compounds **50**–**53** (changes in molecule structure shown in red and green) (*b*); structure of the modelled PcOSBP-oxatiapiprolin complex (red colour corresponds to higher pLDDT values) (*c*); docking results of two isomers of compound **53** (R and S configurations shown in pale blue and pale yellow, hydrogen bonds are highlighted with red dashed lines) (*d*). The figure is published under the CC BY-NC-ND 4.0 license.

It should be noted that the authors [147] conducted field studies with compound **53**, which demonstrated its high antifungal efficacy. This work relates to agrochemistry; however, the approaches to studying the mechanism of action and initial optimization of the structure of primary hit molecules are similar to those used in medicinal chemistry. In the described publication, the AlphaFold model allowed the authors to predict the possible binding mechanism of new molecules and analyze the structure–activity relationship. At the same time, the model was optimized using the MD method before the experiment, which essentially does not allow assessing its initial quality and efficiency for VS purposes. Likely, using a similar approach and a simple homologous model built for the target protein, similar results could have been achieved. The more so, because in the supplementary materials, the authors noted a high overall three-dimensional homology both between the model and 1ZHY (RMSD = 0.54 Å) and in the pocket area (RMSD = 0.65 Å).

5. AlphaFold-latest: a new AlphaFold version

In late October 2023, there was an announcement about the successful work on a new version of the AlphaFold-latest program.[n] Specifically, validation results were presented for the module predicting the spatial position of ligands (small molecules) for various docking methods, including AutoDockVina, Gold, and DiffDock. It was shown that AlphaFold outperforms many docking modelling algorithms in prediction accuracy using the PoseBusters validation sample.[150] The test examples used by the authors were crystallographic complexes reported from May 2022 to January 2023, which were not included in the training set (a total of 8856 complexes after filtering procedures), while the model itself was trained on examples published before September 2019. The spatial overlay of pockets for RMSD calculation was done using protein atoms within 10 Å of ligand atoms. It was noted, in particular, that some problems with the correct placement of ligands in protein models were solved in the new version of the program using docking algorithms discussed in the literature.[112,151] However, the authors did not present results of comparison with the most commonly used commercial programs Glide and MOE (Fig. 17*a*). The results of predicting protein–protein interaction in AlphaFold-latest, particularly for antibody–antigen pairs compared to the previous version (AlphaFold-2.3), are shown in Fig. 17*b*. The accuracy of reproducing positions of nucleic acids and their three-dimensional structures, as well as comparison results with other methods, are reflected in Fig. 17*c*. It is evident that AlphaFold-latest outperforms RoseTTAFold2NA [152] in accuracy but is slightly inferior to Alchemy_RNA2 (under CASP-15 conditions).[153,154] The results of reproducing covalent modifications are shown in Fig. 17*d*. Examples demonstrating that in some cases, the AlphaFold-latest algorithm can predict the true pose of a ligand with good accuracy, unlike docking algorithms, are given in Fig. 18, and the results of validating AlphaFold-latest using various complexes are shown in Fig. 19.

For predicting protein–protein interactions and ligand–protein complex structures, test examples were used that do not have direct analogues in the training data. In the case of protein–protein interactions, only those protein pairs were selected for which there were no training examples with a

[n] https://deepmind.google/discover/blog/a-glimpse-of-the-next-generation-of-alphafold/ (access 28.03.2024).

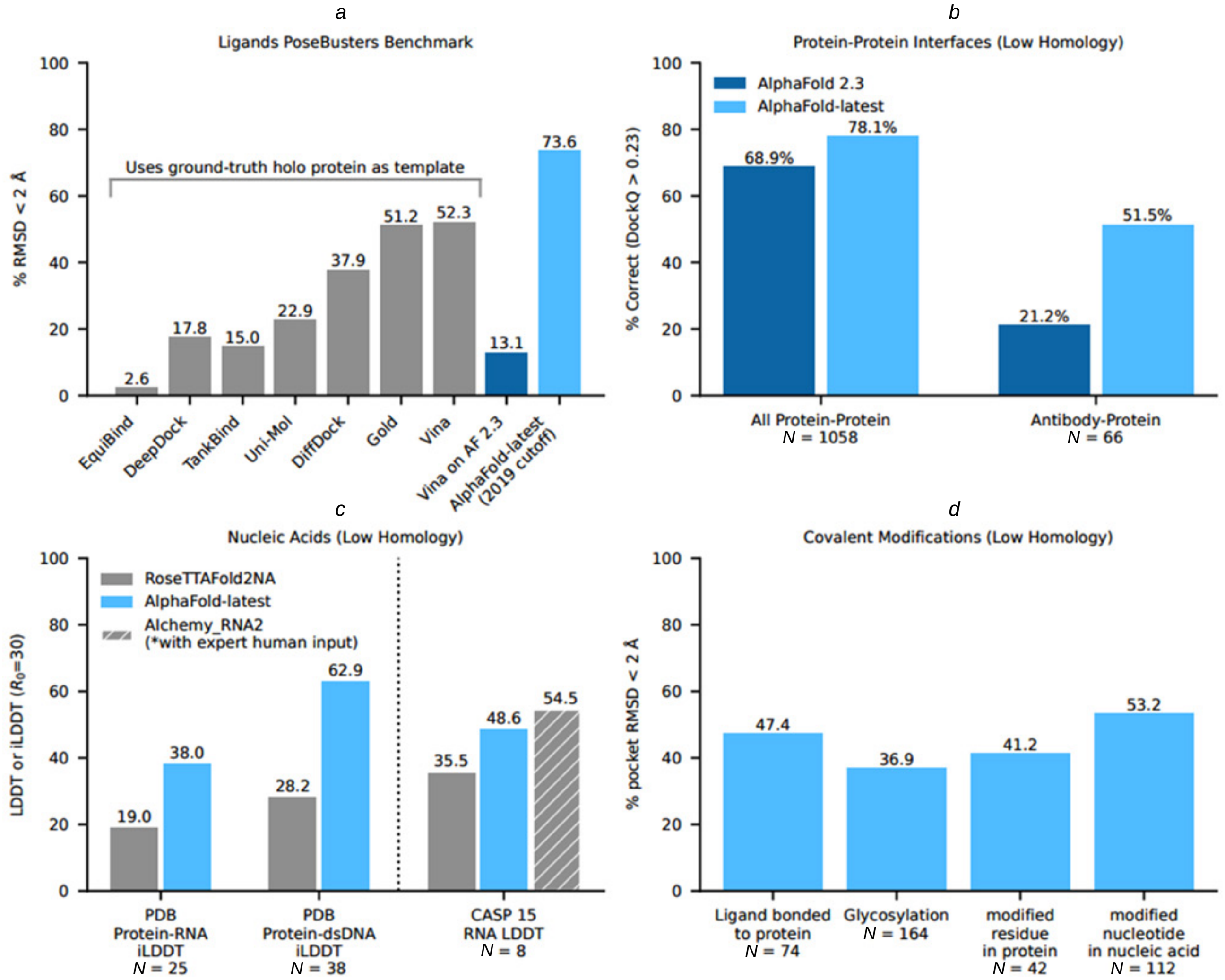
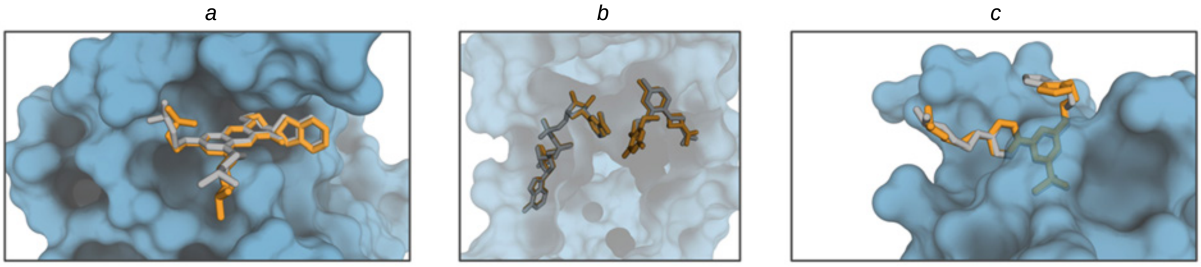


Figure 17. Validation results of AlphaFold-latest models. (*a*) Ligand positioning (428 complexes), (*b*) protein–protein interaction, (*c*) reproduction of the spatial position of nucleic acid within a radius of inclusion of 30 (R_0), and (*d*) accuracy in reproducing covalent ligand poses.[o]

[o] https://deepmind.google/discover/blog/a-glimpse-of-the-next-generation-of-alphafold/ (access 28.03.2024).



PDB ID 7OCB
Best docking RMSD = 4.6 Å
AlphaFold-latest RMSD = 0.96 Å

PDB ID 5SD5
Best docking RMSD = 4.5 Å
AlphaFold-latest RMSD = 0.92 Å

PDB ID 7BLA
Best docking RMSD = 6.3 Å
AlphaFold-latest RMSD = 2.0 Å

Figure 18. Examples of results from the AlphaFold-latest algorithm for ligand–protein complexes where correct poses could not be obtained using Vina and Gold.[p]

[p] https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/a-glimpse-of-the-next-generation-of-alphafold/alphafold_latest_oct2023.pdf (access 28.03.2024).

homology greater than 40% (for the pair). For ligand–protein complexes (small molecules and glycans), the same homology threshold (40%) was used, combined with a Tanimoto coefficient of structural similarity for the molecules not exceeding 0.5. However, information on the homology of pockets, which could significantly contribute to predicting the resulting poses of small molecules, was not provided.

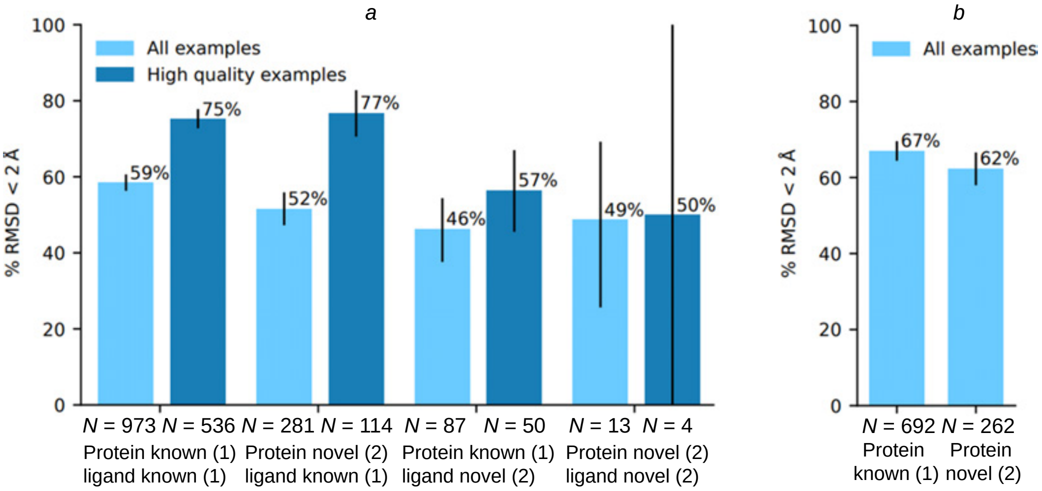


Figure 19. Validation of the AlphaFold-latest algorithm using ligands from the PDB database. (*a*) Small molecules, (*b*) ions; (*1*) complexes with homology to training examples >40% for proteins and Tanimoto coefficient >0.54 (2048 RDKit fingerprints) for ligands, (*2*) for examples in which hese parameters are <40% and <0.5, respectively (the average values for clusters are presented).[r]

[r] https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/a-glimpse-of-the-next-generation-of-alphafold/alphafold_latest_oct2023.pdf (access 28.03.2024).

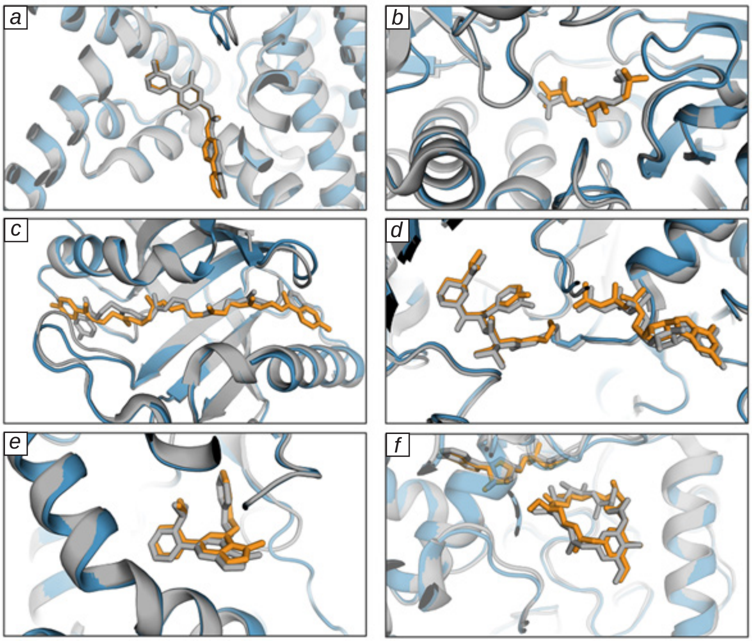


Figure 20. Results of predicting the poses of therapeutic agents in their respective pockets. AlphaFold-latest protein models are highlighted in blue, small molecule poses are light brown, and crystallographic data are in grey. (*a*) LGK974 in the PORCN-WNT3A binding site (PDB: 7URD, RMSD = 0.39 Å); (*b*) (*2S*,*5S*,*6S*)-2,6-bis(azanyl)-5-oxidanyl-7-sulfooxyheptanoic acid in complex with AziU3/U2 (PDB: 7WUX, RMSD = 1.19 Å); (*c*) closthioamide in complex with CtaZ (PDB: 7ZHD, RMSD = 2.22 Å); (*d*) sanguinarine-A analogue covalently bound to KRASG12C in complex with immunophilin CYPA (PDB: 8G9Q, RMSD = 0.80 Å, the covalent bond is not defined); (*e*) NIH-12848 analogue in the allosteric site of PI5P4Kg (PDB: 7QIE, RMSD = 0.85 Å); (*f*) GdmN in complex with macrocyclic 20-O-methyl-19-chloropro-ansamitocin and cofactor (PDB: 7VZN, RMSD = 1.02 Å).[s]

[s] https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/a-glimpse-of-the-next-generation-of-alphafold/alphafold_latest_oct2023.pdf (access 28.03.2024).

For the second validation set, crystallographic data on complexes of proteins with ligands were used, excluding glycans, covalent ligands, ions, and other molecules not related to therapeutic agents. The experiment results are shown in Fig. 20, indicating that the algorithm can predict the poses of ligands for various proteins with comparatively high accuracy. It is clear that, in the foreseeable future, results with biological validation of the described approaches will emerge; such information is necessary to assess the effectiveness of predicting the affinity of small drug molecules.

6. Conclusion

Considering numerous scientific publications in high-impact journals and the results of regular CASP competitions, the AlphaFold algorithm can be considered a leader in the field of modelling three-dimensional structures of protein molecules. This algorithm is crucial for solving many bioinformatics tasks, including analyzing enzymatic reaction mechanisms and self-regulation, profiling mutagenesis sites, modelling protein–protein interactions, and processing X-ray crystallography data. However, most authors of the publications reviewed above agree that without preliminary preparation, AlphaFold models are limitedly suitable for medicinal chemists, especially for docking within typical VS. For example, based on the analysis of experimental results on reproducing true ligand poses,[98] it was concluded that AlphaFold models are mostly unsuitable for VS; it was also noted that due to the peculiarities of training, such models resemble *apo*-forms of proteins, unlike *holo*-variants, which showed better results during docking simulations.

Wong *et al.*[105] concluded that using unmodified AlphaFold models for high-throughput virtual screening is ineffective due to their low classifying ability. The models showed results comparable to those obtained with a random set of molecules at the biological screening stage. Scardino *et al.*[114] also pointed out the unsuitability of AlphaFold models for dockingsimulations. Better results were obtained using experimentally determined spatial structures of proteins, and this trend is characteristic of different docking methods. The Heckelmann's research group [116] concluded that AlphaFold models are not designed to predict the positions of molecules unrelated to the peptide backbone structure, including cofactors, non-protein coenzymes, and small molecules. Other studies reviewed in this overview demonstrated that even when the overall homology of the modelled site does not significantly differ from the experimentally established one, it is possible to obtain positions significantly different from the true ones. Clearly, depending on the conditions in protein molecules, atoms do not occupy static positions, and the conformation of the protein capable of binding a ligand is realized in a single or a limited number of cases, which, in particular, was noted by Nussinov *et al.*,[129,130] who compared AlphaFold models to photos.

Based on the above, it can be asserted that AlphaFold models predict the positions of the peptide backbone atoms and the overall packing with relatively high accuracy, including those for protein molecules for which no close homologues are available. The same cannot be said about the positions of the amino acid side chain atoms lining the pocket, and this fact is critical for modelling the interaction of the ligand with the protein target. Despite this, in experiments for predicting the possible binding site and its geometry, AlphaFold models show better results than homologous models.[151] However, replacing just one amino acid in the binding site or incorrectly modelled positioning of its atoms can lead to a complete loss of ligand affinity. In some cases, irregular protein fragments are located in the binding sites of AlphaFold models, which almost entirely preclude the possibility of obtaining adequate docking simulation results. It must be noted that many publications do not provide results of biological testing of selected molecules, which does not allow for a full assessment of the applicability limits of such models in the early stages of new drug development.

Frequently, the use of standard homologous models for proteins with a high degree of homology in the binding site area leads to comparable and even better modelling results, as pointed out by Karelina *et al.*[151] Many researchers believe that the optimization of AlphaFold models in complex with ligands using molecular dynamics methods or in a flexible docking mode, where the positions of both ligand atoms and amino acids are refined in the force field gradient, can lead to computational models with higher predictive capability compared to unmodified AlphaFold models. However, such approaches require significant time and computational resources and also do not guarantee that the complex chosen as a result of modelling would adequately reflect the molecular mechanism of binding.

It is particularly important to note that for a medicinal chemist, the only and absolute measure of effectiveness for any computational approach is exclusively the results of biological validation. In this context, methods and algorithms are being developed that can automatically adapt AlphaFold structures for high-throughput docking. Specifically, we have proposed an algorithm called AlphaFoldOptimizer, based on machine learning methods, which allows predicting optimal positions of protein atoms in the binding site area and provides optimized three-dimensional models suitable for correct dockingsimulations (a corresponding publication is currently being prepared).

Despite the obvious drawbacks of AlphaFold models, there is no doubt about the rationale for using them in the development of new small medicinal molecules. Many researchers agree that the near future will see modified algorithms based on such models, for example, an improved version of AlphaFold (see Section 5), adapted for effective VS tasks, enriching the arsenal of the modern medicinal chemist with new useful tools.

7. List of abbreviations and designations

ABL1 — Abelson murine leukemia viral oncogene homolog-1 (ABL1 tyrosine kinase),

ADRB2 — β2 adrenergic receptor,

AMPK — 5′-AMP-activated protein kinases (serine/threonine AMP-activated protein kinase),

ANDR — androgen receptor,

ATP — adenosine triphosphate,

auPRC — area under the precision–recall curve (precision and recall metric),

auROC — area under the ROC curve,

AziU3/U2 — aziridine synthase (*Streptomyces sahachiroi*),

BRAF — B-Raf proto-oncogene (serine/threonine kinase),

CASP — critical assessment of protein structure prediction,

CDK20 — cyclin-dependent kinase-20,

CDK2 — cyclin-dependent kinase-2,

CtaZ — GyrI-like protein polythioamide-binding protein,

CYPA — peptidyl-prolyl *cis–trans* isomerase A,

COX1 — cyclooxygenase-1,

DFT-D — dispersion-corrected density functional theory,

DFG — asp(D)-Phe(F)-Gly(G) (a conservative segment in the structure of kinase pockets),

DGK — diacylglycerol kinases,

DGPB — predicted binding energy,

DOOP — docking decoy-based optimized potential,

DRD3 — dopamine receptor D3,

EF — average enrichment factor,

EFS — evaluation function score,

EGFR — epidermal growth factor receptor,

ESR1 — estrogen receptor,

FABP4 — adipocyte fatty acid-binding protein-4,

FA7 — coagulation factor VII,

GdmN — carbamoyltransferase,

GPCR — G protein-coupled receptors,

HDAC — histone deacetylase-3,

HIF-1α — hypoxia-inducible factor 1α,

HSP90 — heat shock protein 90 homologue,

HXK4 — hexokinase-4,

IC_{50} — half-maximal inhibitory concentration,

IGF1R — insulin-like growth factor 1 receptor,

ITAL — integrin α-L,

JMJD8 — JmjC domain-containing protein 8,

KDM — lysine-specific demethylase,

KES1 — oxysterol-binding protein homologue-4,

KPCB — protein kinase C beta-type,

KITH — cytosolic thymidine kinase,

KRAS G12C — GTPase KRas (G12C mutation),

LBDD — ligand-based drug design,

LEV — local environment validation,

LFA1 — lymphocyte function-associated antigen-1,

LOMETS — Local meta-threading server,

MD — molecular dynamics,

MM-GBSA — molecular mechanics with generalized Born and surface area,

MOE — molecular operating environment,

MW — molecular weight,

NAD — nicotinamide adenine dinucleotide,

OfHex1 — chitinolytic β-N-acetyl-D-hexosaminidase,

PDB — protein data bank,

PDE5A — cGMP-specific 3',5'-cyclic phosphodiesterase,

PcOSBP — oxysterol binding protein, PKM2 – Pyruvate kinase type-2,

pLDDT — predicted local distance difference test,

PORCN-WNT3A — the complex of porcupine O-acyltransferase and Wnt family member-3A,

PI5P4Kg — phosphatidylinositol 5-phosphate 4-kinase-g,

PRGR — progesterone receptor,

PTN1 — phosphatidylinositol-3,4,5-trisphosphate 3-phosphatase,

PYRD — mitochondrial dihydroorotate dehydrogenase,

pVHL — Von Hippel-Lindau disease tumour suppressor,

PNPH — purine nucleoside phosphorylase,

Q^3 — a calculated parameter reflecting the accuracy of secondary protein structure prediction,

R^2 — coefficient of determination,

RXRA — retinoic acid receptor-A,

SBDD — structure-based drug design,

SI — selectivity index,

SIK2 — salt inducible kinase-2,

UROK — urokinase-type plasminogen activator,

VS — virtual screening,

τ and τ_0 — intermediate values of torsion angles,

φ and ψ — torsion angles.

8. References

1. A.Grover. *Drug Design: Principles and Applications*. (Springer, 2017); https://doi.org/10.1007/978-981-10-5187-6
2. S.Mandal, M.Moudgil, S.K.Mandal. *Eur. J. Pharmacol.*, **625**, 90 (2009); https://doi.org/10.1016/j.ejphar.2009.06.065
3. R.Baron. *Computational Drug Discovery and Design*. (Totowa, NJ: Humana Press, 2012); https://doi.org/10.1007/978-1-61779-465-0
4. W.P.Janzen. *High Throughput Screening: Methods and Protocols*. (3rd Edn). (New York: Humana Press, 2018); https://doi.org/10.1007/978-1-4939-3673-1
5. D.A.Smith, C.Allerton, A.S.Kalgutkar, H.van de Waterbeemd, D.K.Walker. *Pharmacokinetics and Metabolism in Drug Design*. (John Wiley & Sons, 2012); https://doi.org/10.1002/9783527645763
6. S.Johansson, A.Thakkar, T.Kogej, E.Bjerrum, S.Genheden, T.Bastys, C.Kannas, A.Schliep, H.Chen, O.Engkvist. *Drug Discov. Today Technol.*, **32–33**, 65 (2019); https://doi.org/10.1016/j.ddtec.2020.06.002
7. K.M.Merz, D.Ringe, C.H.Reynolds. *Drug Design: Structure- and Ligand-Based Approaches*. (Cambridge University Press, 2010); https://doi.org/10.1017/CBO9780511730412
8. S.K.Burley, H.M.Berman, G.J.Kleywegt, J.L.Markley, H.Nakamura, S.Velankar. In *Protein Crystallography. Methods in Molecular Biology.* (Eds A.Wlodawer, Z.Dauter, M.Jaskolski). (New York: Humana Press, 2017). Vol. 1607. P. 627; https://doi.org/10.1007/978-1-4939-7000-1_26
9. I.I.Baskin, V.A.Palyulin, N.S.Zefirov. *Russ. Chem. Rev.*, **78**, 495 (2009); https://doi.org/10.1070/RC2009v078n06ABEH004032
10. V.A.Jisna, P.B.Jayaraj. *Protein J.*, **40**, 522 (2021); https://doi.org/10.1007/s10930-021-10003-y
11. M.AlQuraishi. *Curr. Opin. Chem. Biol.*, **65**, 1 (2021); https://doi.org/10.1016/j.cbpa.2021.04.005
12. R.Han, H.Yoon, G.Kim, H.Lee, Y.Lee. *Pharm. Basel Switz.*, **16**, 1259 (2023); https://doi.org/10.3390/ph16091259
13. G.Casadevall, C.Duran, S.Osuna. *JACS Au*, **3**, 1554 (2023); https://doi.org/10.1021/jacsau.3c0018
14. M.A.Pak, K.A.Markhieva, M.S.Novikova, D.S.Petrov, I.S.Vorobyev, E.S.Maksimova, F.A.Kondrashov, D.N.Ivankov. *PLOS ONE*, **18**, e0282689 (2023); https://doi.org/10.1371/journal.pone.0282689
15. A.Harmalkar, S.Lyskov, J.J.Gray. *BioRxivPrepr. Serv. Biol.*, 2023.07.28.551063 (2023); https://doi.org/10.1101/2023.07.28.551063
16. S.E.Biehn, S.Lindert. *Annu. Rev. Phys. Chem.*, **73**, 1 (2022); https://doi.org/10.1146/annurev-physchem-082720-123928
17. D.V.Laurents. *Front. Mol. Biosci.*, **9**, 906437 (2022); https://doi.org/10.3389/fmolb.2022.906437
18. J.J.Clark, M.L.Benson, R.D.Smith, H.A.Carlson. *PLOS Comput. Biol.*, **15**, e1006705 (2019); https://doi.org/10.1371/journal.pcbi.1006705
19. J.E.Ladbury. *Chem. Biol.*, **3**, 973 (1996); https://doi.org/10.1016/S1074-5521(96)90164-7
20. C.S.Poornima, P.M.Dean. *J. Comput. Aided Mol. Des.*, **9**, 500 (1995); https://doi.org/10.1007/BF00124321
21. M.H.Daniels, G.Malojcic, S.L.Clugston, B.Williams, M.Coeffet-Le Gal, X.-R.Pan-Zhou, S.Venkatachalan, J.-C.Harmange, M.Ledeboer. *J. Med. Chem.*, **65**, 3575 (2022); https://doi.org/10.1021/acs.jmedchem.1c02069
22. J.-Y.Zhu, R.A.Cuellar, N.Berndt, H.E.Lee, S.H.Olesen, M.P.Martin, J.T.Jensen, G.I.Georg, E.Schönbrunn. *J. Med. Chem.*, **60**, 7863 (2017); https://doi.org/10.1021/acs.jmedchem.7b00996
23. R.Thaimattam, R.Banerjee, R.Miglani, J.Iqbal. *Curr. Pharm. Des.*, **13**, 2751 (2007); https://doi.org/10.2174/138161207781757042
24. L.Zara, N.-L.Efrém, J.E.Van Muijlwijk-Koezen, I.J.P.De Esch, B.Zarzycka. *Drug Discov. Today: Technol.*, **40**, 36 (2021); https://doi.org/10.1016/j.ddtec.2021.10.001
25. V.B.Luzhkov. *Russ. Chem. Rev.*, **86**, 211 (2017); https://doi.org/10.1070/RCR4610
26. C.B.Anfinsen. *Science*, **181**, 223 (1973); https://doi.org/10.1126/science.181.4096.223
27. A.Deiana, S.Forcelloni, A.Porrello, A.Giansanti. *PlOS ONE*, **14**, e0217889 (2019); https://doi.org/10.1371/journal.pone.0217889
28. P.E.Wright, H.J.Dyson. *Curr. Opin. Struct. Biol.*, **19**, 31 (2009); https://doi.org/10.1016/j.sbi.2008.12.003
29. J.S.Richardson. In *Advances in Protein Chemystry.* Vol. 34. (Elsevier, 1981). P. 167; https://doi.org/10.1016/S0065-3233(08)60520-3
30. A.Šali, T.L.Blundell. *J. Mol. Biol.*, **234**, 779 (1993); https://doi.org/10.1006/jmbi.1993.1626
31. A.Waterhouse, M.Bertoni, S.Bienert, G.Studer, G.Tauriello, R.Gumienny, F.T.Heer, T.A.P.de Beer, C.Rempfer, L.Bordoli, R.Lepore, T.Schwede. *Nucleic Acids Res.*, **46**, W296 (2018); https://doi.org/10.1093/nar/gky427
32. S.B.Needleman, C.D.Wunsch. *J. Mol. Biol.*, **48**, 443 (1970); https://doi.org/10.1016/0022-2836(70)90057-4
33. T.F.Smith, M.S.Waterman. *J. Mol. Biol.*, **147**, 195 (1981); https://doi.org/10.1016/0022-2836(70)90057-4
34. S.F.Altschul, W.Gish, W.Miller, E.W.Myers, D.J.Lipman. *J. Mol. Biol.*, **215**, 403 (1990); https://doi.org/10.1016/S0022-2836(05)80360-2
35. J.Söding. *Bioinformatics*, **21**, 951 (2005); https://doi.org/10.1093/bioinformatics/bti125
36. Y.Yang, E.Faraggi, H.Zhao, Y.Zhou. *Bioinformatics*, **27**, 2076 (2011); https://doi.org/10.1093/bioinformatics/btr350
37. K.Ginalski, A.Elofsson, D.Fischer, L Rychlewski. *Bioinformatics*, **19**, 1015 (2003); https://doi.org/10.1093/bioinformatics/btr350
38. S.Wu, Y.Zhang. *Nucleic Acids Res.*, **35**, 3375 (2007); https://doi.org/10.1093/nar/gkm251
39. A.Kryshtafovych, B.Monastyrskyy, K.Fidelis, J.Moult, T.Schwede, A.Tramontano. *Proteins Struct. Funct. Bioinform.*, **86**, 321 (2018); https://doi.org/10.1002/prot.25425

40. V.Modi, Q.Xu, S.Adhikari, R.L.Dunbrack. *Proteins Struct. Funct. Bioinform.*, **84**, 200 (2016); https://doi.org/10.1002/prot.25049
41. J.Skolnick, H.Zhou. *J. Phys. Chem. B*, **121**, 3546 (2017); https://doi.org/10.1021/acs.jpcb.6b09517
42. A.Kryshtafovych, T.Schwede, M.Topf, K.Fidelis, J.Moult. *Proteins Struct. Funct. Bioinform.*, **87**, 1011 (2019); https://doi.org/10.1002/prot.25823
43. A.T.Brünger, P.D.Adams, G.M.Clore, W.L.DeLano, P.Gros, R.W.Grosse-Kunstleve, J.S.Jiang, J.Kuszewski, M.Nilges, N.S.Pannu, R.J.Read, L.M.Rice, T.Simonson, G.L.Warren. *Acta Cryst.*, **D54**, 905 (1998); https://doi.org/10.1107/S0907444998003254
44. C.A.Rohl, C.E.M.Strauss, K.M.S.Misura, D.Baker. In *Methods Enzymology.* Vol. 383. (Elsevier, 2004). P. 66; https://doi.org/10.1016/S0076-6879(04)83004-0
45. J.Soding, A.Biegert, A.N.Lupas. *Nucleic Acids Res.*, **33**, W244 (2005); https://doi.org/10.1093/nar/gki408
46. B.Webb, A.Sali. *Curr. Protoc. Bioinform.*, **54**, (2016); https://doi.org/10.1002/cpbi.3
47. J.Lundström, L.Rychlewski, J.Bujnicki, A.Elofsson. *Protein Sci.*, **10**, 2354 (2001); https://doi.org/10.1110/ps.08501
48. J.Yang, R.Yan, A.Roy, D.Xu, J.Poisson, Y.Zhang. *Nat. Methods*, **12**, 7 (2015); https://doi.org/10.1038/nmeth.3213
49. L.A.Kelley, S.Mezulis, C.M.Yates, M.N.Wass, M.J.E.Sternberg. *Nat. Protoc.*, **10**, 845 (2015); https://doi.org/10.1038/nprot.2015.053
50. S.M.Mortuza, W.Zheng, C.Zhang, Y.Li, R.Pearce, Y.Zhang. *Nat. Commun.*, **12**, 5011 (2021); https://doi.org/10.1038/s41467-021-25316-w
51. M.Torrisi, G.Pollastri, Q.Le. *Comput. Struct. Biotechnol. J.*, **18**, 1301 (2020); https://doi.org/10.1016/j.csbj.2019.12.011
52. J.Jumper, R.Evans, A.Pritzel, T.Green, M.Figurnov, O.Ronneberger, K.Tunyasuvunakool, R.Bates, A.Žídek, A.Potapenko, A.Bridgland, C.Meyer, S.A.A.Kohl, A.J.Ballard, A.Cowie, B.Romera-Paredes, S.Nikolov, R.Jain, J.Adler, T.Back, S.Petersen, D.Reiman, E.Clancy, M.Zielinski, M.Steinegger, M.Pacholska, T.Berghammer, S.Bodenstein, D.Silver, O.Vinyals, A.W.Senior, K.Kavukcuoglu, P.Kohli, D.Hassabis. *Nature*, **596**, 583 (2021); https://doi.org/10.1038/s41586-021-03819-2
53. R.Wu, F.Ding, R.Wang, R.Shen, X.Zhang, S.Luo, C.Su, Z.Wu, Q.Xie, B.Berger, J.Ma, J.Peng. *Bioinformatics*, 2022; https://doi.org/10.1101/2022.07.21.500999
54. Z.Lin, H.Akin, R.Rao, B.Hie, Z.Zhu, W.Lu, N.Smetanin, R.Verkuil, O.Kabeli, Y.Shmueli, A.Dos Santos Costa, M.Fazel-Zarandi, T.Sercu, S.Candido, A.Rives. *Science*, **379**, 1123 (2023); https://doi.org/10.1126/science.ade2574
55. N.Guex, M.C.Peitsch, T.Schwede. *Electrophoresis*, **30**, (2009); https://doi.org/10.1002/elps.200900140
56. Y.Zhang, J.Skolnick. *Proc. Natl. Acad. Sci.*, **101**, 7594 (2004); https://doi.org/10.1073/pnas.0305695101
57. M.-H.Chae, F.Krull, E.W.Knapp. *Proteins Struct. Funct. Bioinform.*, **83**, 881 (2015); https://doi.org/10.1002/prot.24782
58. Y.Zhang, J.Skolnick. *J. Comput. Chem.*, **25**, 865 (2004); https://doi.org/10.1002/jcc.20011
59. Y.Zhang. *Nucleic Acids Res.*, **33**, 2302 (2005); https://doi.org/10.1093/nar/gki524
60. Y.Li, Y.Zhang. *Proteins Struct. Funct. Bioinform.*, **76**, 665 (2009); https://doi.org/10.1002/prot.22380
61. Y.Ali, M.Kausar, M.Farooq, N.Farooqi, Z.Ul Islam, S.Khan, A.Aman, N.Khan, A.Kamil, F.Jalil. *PLOS ONE*, **18**, e0285874 (2023); https://doi.org/10.1371/journal.pone.0285874
62. H.Mani, C.C.Chang, H.-J.Hsu, C.-H.Yang, J.-H.Yen, J.-W.Liou. *Bioengineering*, **10**, 1004 (2023); https://doi.org/10.3390/bioengineering10091004
63. W.Zheng, Q.Wuyun, P.L.Freddolino, Y.Zhang. *Proteins Struct. Funct. Bioinform.*, **91**, 1684 (2023); https://doi.org/10.1002/prot.26585
64. G.Senthil Kumar, N.Kishore, E.Elumalai, K.K.Gupta. *J. Biomol. Struct. Dyn.*, 1 (2023); https://doi.org/10.1080/07391102.2023.2246575
65. X.Huang, R.Pearce, G.S.Omenn, Y.Zhang. *Int. J. Mol. Sci.*, **22**, 7060 (2021); https://doi.org/10.3390/ijms22137060
66. J.Abbass, J.C.Nebel. *Curr. Bioinform.*, **15**, 611 (2020); https://doi.org/10.2174/1574893615999200504103643
67. S.Altschul. *Nucleic Acids Res.*, **25**, 3389 (1997); https://doi.org/10.1093/nar/25.17.3389
68. A.Leaver-Fay, M.Tyka, S.M.Lewis, O.F.Lange, J.Thompson, R.Jacak, K.W.Kaufman, P.D.Renfrew, C.A.Smith, W.Sheffler, I.W.Davis, S.Cooper, A.Treuille, D.J.Mandell, F.Richter, Y.-E.A.Ban, S.J.Fleishman, J.E.Corn, D.E.Kim, S.Lyskov, M.Berrondo, S.Mentzer, Z.Popović, J.J.Havranek, J.Karanicolas, R. Das, J. Meiler, T. Kortemme, J.J.Gray, B.Kuhlman, D.Baker, P.Bradley. In *Methods in Enzymology.* (Elsevier, 2011). P. 545; https://doi.org/10.1016/B978-0-12-381270-4.00019-6
69. P.Bradley, K.M.S.Misura, D.Baker. *Science*, **309**, 1868 (2005); https://doi.org/10.1126/science.1113801
70. D.Röthlisberger, O.Khersonsky, A.M.Wollacott, L.Jiang, J.DeChancie, J.Betker, J.L.Gallaher, E.A.Althoff, A.Zanghellini, O.Dym, S.Albeck, K.N.Houk, D.S.Tawfik, D.Baker. *Nature*, **453**, 190 (2008); https://doi.org/10.1038/nature06879
71. S.Cooper, F.Khatib, A.Treuille, J.Barbero, J.Lee, M.Beenen, A.Leaver-Fay, D.Baker, Z.Popović, F.Players. *Nature*, **466**, 756 (2010); https://doi.org/10.1038/nature09304
72. J.B.Siegel, A.L.Smith, S.Poust, A.J.Wargacki, A.Bar-Even, C.Louw, B.W.Shen, C.B.Eiben, H.M.Tran, E.Noor, J.L.Gallaher, J.Bale, Y.Yoshikuni, M.H.Gelb, J.D.Keasling, B.L.Stoddard, M.E.Lidstrom, D.Baker. *Proc. Natl. Acad. Sci.*, **112**, 3704 (2015); https://doi.org/10.1073/pnas.1500545112
73. H.Zhou, Y.Zhou. *Proteins Struct. Funct. Bioinform.*, **55**, 1005 (2004); https://doi.org/10.1002/prot.20007
74. W.Zhang, A.K.Dunker, Y.Zhou. *Proteins Struct. Funct. Bioinform.*, **71**, 61 (2008); https://doi.org/10.1002/prot.21654
75. H.Zhou, Y.Zhou. *Proteins Struct. Funct. Bioinform.*, **58**, 321 (2005); https://doi.org/10.1002/prot.20308
76. S.Liu, C.Zhang, S.Liang, Y.Zhou. *Proteins Struct. Funct. Bioinform.*, **68**, 636 (2007); https://doi.org/10.1002/prot.21459
77. W.Zhang, S.Liu, Y.Zhou. *PLOS ONE*, **3**, e2325 (2008); https://doi.org/10.1371/journal.pone.0002325
78. H.Zhou, Y.Zhou. *Proteins Struct. Funct. Bioinform.*, **61**, 152 (2005); https://doi.org/10.1002/prot.20732
79. J.N.D.Battey, J.Kopp, L.Bordoli, R.J.Read, N.D.Clarke, T.Schwede. *Proteins Struct. Funct. Bioinform.*, **69**, 68 (2007); https://doi.org/10.1002/prot.21761
80. J.Hargbo, A.Elofsson. *Proteins*, **36**, 68 (1999)
81. B.Rost, C.Sander, R.Schneider. *J. Mol. Biol.*, **235**, 13 (1994); https://doi.org/10.1016/S0022-2836(05)80007-5
82. N.Siew, A.Elofsson, L.Rychlewski, D.Fischer. *Bioinformatics*, **16**, 776 (2000); https://doi.org/10.1093/bioinformatics/16.9.776
83. J.Meiler, M.Müller, A.Zeidler, F.Schmäschke. *J. Mol. Model.*, **7**, 360 (2001); https://doi.org/10.1007/s008940100038
84. E.Faraggi, T.Zhang, Y.Yang, L.Kurgan, Y.Zhou. *J. Comput. Chem.*, **33**, 259 (2012); https://doi.org/10.1002/jcc.21968
85. O.Dor, Y.Zhou. *Proteins Struct. Funct. Bioinform.*, **68**, 76 (2007); https://doi.org/10.1002/prot.21408
86. G.Wang, R.L.Dunbrack. *Bioinformatics*, **19**, 1589 (2003); https://doi.org/10.1093/bioinformatics/btg224
87. A.Natarajan, R.Thangarajan, S.Kesavan. *Asian Pac. J. Cancer Prev.*, **20**, 3399 (2019); https://doi.org/10.31557/APJCP.2019.20.11.3399
88. M.A.Mia, M.N.Uddin, Y.Akter, Jesmin, L.Wal Marzan. *Bioinform. Biol. Insights*, **16**, (2022); https://doi.org/10.1177/11779322221104308
89. H.J.Tey, C.H.Ng. *PeerJ*, **7**, e7667 (2019); https://doi.org/10.7717/peerj.7667

90. S.Geethu, E.R.Vimina. *Protein J.*, **40**, 669 (2021); https://doi.org/10.1007/s10930-021-10016-7
91. A.Prakash, M.Jeffryes, A.Bateman, R.D.Finn. *Curr. Protoc. Bioinform.*, **60** (2017); https://doi.org/10.1002/cpbi.40
92. M.Steinegger, M.Meier, M.Mirdita, H.Vöhringer, S.J.Haunsberger, J.Söding. *BMC Bioinformatics*, **20**, 473 (2019); https://doi.org/10.1186/s12859-019-3019-7
93. Q.Xie, M.-T.Luong, E.Hovy, Q.V.Le. *Self-training with Noisy Student Improves ImageNet Classification*. arXiv:1911.04252v4 (2019); https://doi.org/10.48550/ARXIV.1911.04252
94. A.V.Finkelstein. *Bioinformatics*, 2022; https://doi.org/10.1101/2022.11.21.517308
95. M.Baek, F.DiMaio, I.Anishchenko, J.Dauparas, S.Ovchinnikov, G.R.Lee, J. Wang, Q.Cong, L.N.Kinch, R.D.Schaeffer, C.Millán, H.Park, C.Adams, C.R.Glassman, A.DeGiovanni, J.H.Pereira, A.V.Rodrigues, A.A.Van Dijk, A.C.Ebrecht, D.J.Opperman, T.Sagmeister, C.Buhlheller, T.Pavkov-Keller, M.K.Rathinaswamy, U.Dalwadi, C.K.Yip, J.E.Burke, K.C.Garcia, N.V.Grishin, P.D.Adams, R.J.Read, D.Baker. *Science*, **373**, 871 (2021); https://doi.org/10.1126/science.abj8754
96. Z.Yang, X.Zeng, Y.Zhao, R.Chen. *Sig. Transduct. Target. Ther.*, **8**, 115 (2023); https://doi.org/10.1038/s41392-023-01381-z
97. T.A.Halgren, R.B.Murphy, R.A.Friesner, H.S.Beard, L.L.Frye, W.T.Pollard, J.L.Banks. *J. Med. Chem.*, **47**, 1750 (2004); https://doi.org/10.1021/jm030644s
98. Y.Zhang, M.Vass, D.Shi, E.Abualrous, J.M.Chambers, N.Chopra, C.Higgs, K.Kasavajhala, H.Li, P.Nandekar, H.Sato, E.B.Miller, M.P.Repasky, S.V.Jerome. *J. Chem. Inf. Model.*, **63**, 1656 (2023); https://doi.org/10.1021/acs.jcim.2c01219
99. S.Chen, Z.Sun, L.Lin, Z.Liu, X.Liu, Y.Chong, Y.Lu, H.Zhao, Y.Yang. *J. Chem. Inf. Model.*, **60**, 391 (2020); https://doi.org/10.1021/acs.jcim.9b00438
100. H.Guterres, S.-J.Park, W.Jiang, W Im. *J. Chem. Inf. Model.*, **61**, 535 (2021); https://doi.org/10.1021/acs.jcim.0c01354
101. C.R.Søndergaard, M.H.M.Olsson, M.Rostkowski, J.H.Jensen. *J. Chem. Theory Comput.*, **7**, 2284 (2011); https://doi.org/10.1021/ct200133y
102. J.-F.Truchon, C.I.Bayly. *J. Chem. Inf. Model.*, **47**, 488 (2007); https://doi.org/10.1021/ci600426e
103. Y.Weng, C.Pan, Z.Shen, S.Chen, L.Xu, X.Dong, J.Chen. *Evid. Based Complement. Alternat. Med.*, **2022**, ID 4629392 (2022); https://doi.org/10.1155/2022/4629392
104. L.Solis-Vasquez, A.F.Tillack, D.Santos-Martins, A.Koch, S.LeGrand, S.Forli. *Parallel Comput.*, **109**, 102861 (2022); https://doi.org/10.1016/j.parco.2021.102861
105. F.Wong, A.Krishnan, E.J.Zheng, H.Stärk, A.L.Manson, A.M.Earl, T.Jaakkola, J.J.Collins. *Mol. Syst. Biol.*, **18**, e11081 (2022); https://doi.org/10.15252/msb.202211081
106. W.J.Allen, T.E.Balius, S.Mukherjee, S.R.Brozell, D.T.Moustakas, P.T.Lang, D.A.Case, I.D.Kuntz, R.C.Rizzo. *J. Comput. Chem.*, **36**, 1132 (2015); https://doi.org/10.1002/jcc.23905
107. P.J.Ballester, J.B.O.Mitchell. *Bioinformatics*, **26**, 1169 (2010); https://doi.org/10.1093/bioinformatics/btq112
108. M.Wójcikowski, P.J.Ballester, P.Siedlecki. *Sci. Rep.*, **7**, 46710 (2017); https://doi.org/10.1038/srep46710
109. M.Wójcikowski, M.Kukiełka, M.M.Stepniewska-Dziubinska, P.Siedlecki. *Bioinformatics*, **35**, 1334 (2019); https://doi.org/10.1093/bioinformatics/bty757
110. J.D.Durrant, J.A.McCammon. *J. Chem. Inf. Model.*, **50**, 1865 (2010); https://doi.org/10.1021/ci100244v
111. A.Mullard. *Nat. Rev. Drug Discov.*, **20**, 725 (2021); https://doi.org/10.1038/d41573-021-00161-0
112. V.Scardino, J.I.Di Filippo, C.N.Cavasotto. *iScience*, **26**, 105920 (2023); https://doi.org/10.1016/j.isci.2022.105920
113. K.Palacio-Rodríguez, I.Lans, C.N.Cavasotto, P.Cossio. *Sci. Rep.*, **9**, 5142 (2019); https://doi.org/10.1038/s41598-019-41594-3
114. V.Scardino, M.Bollini, C.N.Cavasotto. *RSC Adv.*, **11**, 35383 (2021); https://doi.org/10.1039/D1RA05785E
115. A.M.Díaz-Rovira, H.Martín, T.Beuming, L.Díaz, V.Guallar, S.S.Ray. *Biochemistry* (2022); https://doi.org/10.1101/2022.08.18.504412
116. M.L.Hekkelman, I.De Vries, R.P.Joosten, A.Perrakis. *Nat. Methods*, **20**, 205 (2023); https://doi.org/10.1038/s41592-022-01685-y
117. J.D.Fischer, G.L.Holliday, J.M.Thornton. *Bioinformatics*, **26**, 2496 (2010); https://doi.org/10.1093/bioinformatics/btq442
118. E.Krieger, G.Vriend. *Bioinformatics*, **30**, 2981 (2014); https://doi.org/10.1093/bioinformatics/btu426
119. X.Liang, H Zhang, Z.Wang, X.Zhang, Z.Dai, J.Zhang, P.Luo, L.Zhang, J.Hu, Z.Liu, C.Bi, Q.Cheng. *Front. Immunol.*, **13**, 875786 (2022); https://doi.org/10.3389/fimmu.2022.875786
120. A.Subramanian, R.Narayan, S.M.Corsello, D.D.Peck, T.E.Natoli, X.Lu, J.Gould, J.F.Davis, A.A.Tubelli, J.K.Asiedu, D.L.Lahr, J.E.Hirschman, Z.Liu, M.Donahue, B.Julian, M.Khan, D.Wadden, I.C.Smith, D.Lam, A.Liberzon, C.Toder, M.Bagul, M.Orzechowski, O.M.Enache, F.Piccioni, S.A.Johnson, N.J.Lyons, A.H.Berger, A.F.Shamji, A.N.Brooks, A.Vrcic, C.Flynn, J.Rosains, D.Y.Takeda, R.Hu, D.Davison, J.Lamb, K.Ardlie, L.Hogstrom, P.Greenside, N.S.Gray, P.A.Clemons, S.Silver, X.Wu, W.-N.Zhao, W.Read-Button, X.Wu, S.J.Haggarty, L.V.Ronco, J.S.Boehm, S.L.Schreiber, J.G.Doench, J.A.Bittker, D.E.Root, B.Wong, T.R.Golub. *Cell*, **171**, 1437 (2017); https://doi.org/10.1016/j.cell.2017.10.049
121. A.T.Satti, A.R.Siddiqi, A.Maryam, S.Chaitanya Vedithi, T.L.Blundell. *J. Biomol. Struct. Dyn.*, 1 (2023); https://doi.org/10.1080/07391102.2023.2264394
122. I.W.Davis, A.Leaver-Fay, V.B.Chen, J.N.Block, G.J.Kapral, X.Wang, L.W.Murray, W.B.Arendall, J.Snoeyink, J.S.Richardson, D.C.Richardson. *Nucleic Acids Res.*, **35**, W375 (2007). https://doi.org/10.1093/nar/gkm216
123. S. Genheden, U. Ryde. *Expert Opin. Drug Discov.*, **10**, 449 (2015); https://doi.org/10.1517/17460441.2015.1032936
124. K.B.Lokhande, A.Shrivastava, A.Singh. *J. Biomol. Struct. Dyn.*, 1 (2023); https://doi.org/10.1080/07391102.2023.2183342
125. S.N.Shchelkunov, A.V.Totmenin, I.V.Babkin, P.F.Safronov, O.I.Ryazankina, N.A.Petrov, V.V.Gutorov, E.A.Uvarova, M.V.Mikheev, J.R.Sisler, J.J.Esposito, P.B.Jahrling, B.Moss, L.S.Sandakhchiev. *FEBS Lett.*, **509**, 66 (2001); https://doi.org/10.1016/S0014-5793(01)03144-1
126. M.Wiederstein, M.J.Sippl. *Nucleic Acids Res.*, **35**, W407 (2007); https://doi.org/10.1093/nar/gkm290
127. H.Y.I.Lam, J.S.Guan, Y.Mu. *Molecules*, **27**, 5277 (2022); https://doi.org/10.3390/molecules27165277
128. B.Delley. *J. Chem. Phys.*, **92**, 508 (1990); https://doi.org/10.1063/1.458452
129. R.Nussinov, M.Zhang, Y.Liu, H.Jang. *Drug Discov. Today*, **28**, 103551 (2023); https://doi.org/10.1016/j.drudis.2023.103551
130. R.Nussinov, M.Zhang, Y.Liu, H Jang. *J. Phys. Chem. B*, **126**, 6372 (2022); https://doi.org/10.1021/acs.jpcb.2c04346
131. D.Réa, T.P.Hughes. *Crit. Rev. Oncol. Hematol.*, **171**, 103580 (2022); https://doi.org/10.1016/j.critrevonc.2022.103580
132. F.-Y.Lin, J.Li, Y.Xie, J.Zhu, T.T.Huong Nguyen, Y.Zhang, J.Zhu, T.A.Springer. *Cell*, **185**, 3533 (2022); https://doi.org/10.1016/j.cell.2022.08.008
133. L.Heo, M.Feig. *Protein. Struct. Funct. Bioinform.*, **90**, 1873 (2022); https://doi.org/10.1002/prot.26382
134. F.Ren, X.Ding, M.Zheng, M.Korzinkin, X.Cai, W.Zhu, A.Mantsyzov, A.Aliper, V.Aladinskiy, Z.Cao, S.Kong, X.Long, B.H.Man Liu, Y.Liu, V.Naumov, A.Shneyderman, I.V.Ozerov, J.Wang, F.W.Pun, D.A.Polykovskiy, C.Sun, M.Levitt, A.Aspuru-Guzik, A.Zhavoronkov. *Chem. Sci.*, **14**, 1443 (2023); https://doi.org/10.1039/D2SC05709C
135. A.Olsen, Z.Harpaz, C.Ren, A.Shneyderman, A.Veviorskiy, M.Dralkina, S.Konnov, O.Shcheglova, F.W.Pun, G.H.D.Leung, H.W.Leung, I.V.Ozerov, A.Aliper,

M.Korzinkin, A.Zhavoronkov. *Aging* (2023); https://doi.org/10.18632/aging.204678

136. Y.A.Ivanenkov, D.Polykovskiy, D.Bezrukov, B.Zagribelnyy, V.Aladinskiy, P.Kamya, A.Aliper, F.Ren, A.Zhavoronkov. *J. Chem. Inf. Model.*, **63**, 695 (2023); https://doi.org/10.1021/acs.jcim.2c01191
137. A.Monastyrskyi, S.Bayle, V.Quereda, W.Grant, M.Cameron, D.Duckett, W.Roush. *Bioorg. Med. Chem. Lett.*, **28**, 400 (2018); https://doi.org/10.1016/j.bmcl.2017.12.026
138. G.Shi, H.Scott, N.I.F.M.Azhar, A.Gialeli, B.Clennell, K.S.Lee, J.Hurcombe, D.Whitcomb, R.Coward, L.-F.Wong, O.Cordero-Llana, J.B.Uney. *Sci. Rep.*, **13**, 8334 (2023); https://doi.org/10.1038/s41598-023-35480-2
139. W.Zhu, X.Liu, Q.Li, F.Gao, T.Liu, X.Chen, M.Zhang, A.Aliper, F.Ren, X.Ding, A.Zhavoronkov. *Bioorg. Med. Chem.*, **91**, 117414 (2023); https://doi.org/10.1016/j.bmc.2023.117414
140. Y.Liu, W.Tang, C.Ji, J.Gu, Y.Chen, J.Huang, X.Zhao, Y.Sun, C.Wang, W.Guan, J.Liu, B.Jiang. *Front. Pharmacol.*, **11**, 624429 (2021); https://doi.org/10.3389/fphar.2020.624429
141. R.Tesch, M.Rak, M.Raab, L.M.Berger, T.Kronenberger, A.C.Joerger, B.-T.Berger, I.Abdi, T.Hanke, A.Poso, K.Strebhardt, M.Sanhaji, S.Knapp. *J. Med. Chem.*, **64**, 8142 (2021); https://doi.org/10.1021/acs.jmedchem.0c02144
142. R.Mendez, M.Shaikh, M.C.Lemke, K.Yuan, A.H.Libby, D.L.Bai, M.M. Ross, T.E.Harris, K.-L.Hsu. *RSC Chem. Biol.*, **4**, 422 (2023); https://doi.org/10.1039/D3CB00057E
143. C.E.Franks, S.T.Campbell, B.W.Purow, T.E.Harris, K.-L.Hsu. *Cell Chem. Biol.*, **24**, 870 (2017); https://doi.org/10.1016/j.chembiol.2017.06.007
144. A.Meller, S.De Oliveira, A.Davtyan, T.Abramyan, G.R.Bowman, H.Van Den Bedem. *Front. Mol. Biosci.*, **10**, 1171143 (2023); https://doi.org/10.3389/fmolb.2023.1171143
145. L.Li, C.Meyer, Z.-W.Zhou, A.Elmezayen, K.Westover. *J. Mol. Biol.*, **434**, 167626 (2022); https://doi.org/10.1016/j.jmb.2022.167626
146. L.Perna, M.Castelli, E.Frasnetti, L.E.L.Romano, G.Colombo, C.Prodromou, J.P.Chapple. *Front. Mol. Biosci.*, **9**, 1074714 (2023); https://doi.org/10.3389/fmolb.2022.1074714
147. J.-L.Li, J.-F.Yang, L.-M.Zhou, M.Cai, Z.-Q.Huang, X.-L.Liu, X.-L.Zhu, G.-F.Yang. *J. Agric. Food Chem.*, **71**, 9519 (2023); https://doi.org/10.1021/acs.jafc.3c00990
148. Z.Wang, H.Sun, X.Yao, D.Li, L.Xu, Y.Li, S.Tian, T.Hou. *Phys. Chem. Chem. Phys.*, **18**, 12964 (2016); https://doi.org/10.1039/C6CP01555G
149. Y.Gao, X.Zhao, X.Sun, Z.Wang, J.Zhang, L.Li, H.Shi, M.Wang. *J. Agric. Food Chem.*, **69**, 3289 (2021); https://doi.org/10.1021/acs.jafc.0c04163
150. M.Buttenschoen, G.M.Morris, C.M.Deane. (2023); https://doi.org/10.48550/ARXIV.2308.05777
151. M.Karelina, J.J.Noh, R.O.Dror. *eLife*, (2023); https://doi.org/10.7554/eLife.89386.1
152. M.Baek, R.McHugh, I.Anishchenko, D.Baker, F.DiMaio. *Bioinformatics* (2022); https://doi.org/10.1101/2022.09.09.507333
153. K.Chen, Y.Zhou, S.Wang, P.Xiong. *Proteins Struct. Funct. Bioinform.*, **91**, 1771 (2023); https://doi.org/10.1002/prot.26574
154. P.Xiong, R.Wu, J.Zhan, Y.Zhou. *Nat. Commun.*, **12**, 2777 (2021); https://doi.org/10.1038/s41467-021-23100-4