Published online 30 April 2022

https://doi.org/10.1093/nar/gkac305

Proteins*Plus*: a comprehensive collection of web-based molecular modeling tools

Katrin Schöning-Stierand[1,†], Konrad Diedrich[1,†], Christiane Ehrt[1,†], Florian Flachsenberg[1,†], Joel Graef[1,†], Jochen Sieg[1,†], Patrick Penner[1], Martin Poppinga[1,2], Annett Ungethüm[3] and Matthias Rarey[1,*]

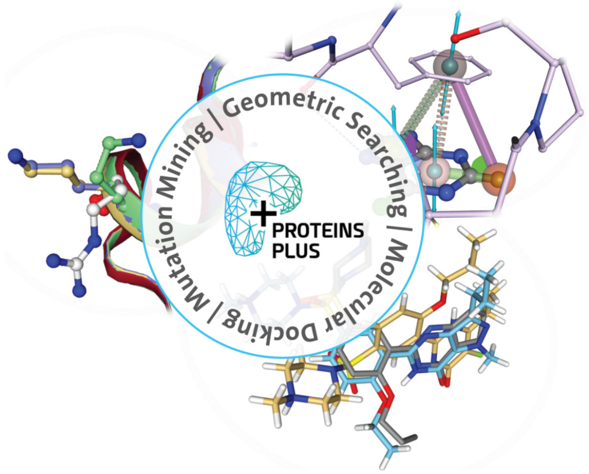
[1]Universität Hamburg, ZBH - Center for Bioinformatics, Bundesstraße 43, 20146 Hamburg, Germany, [2]Universität Hamburg, Department of Informatics, Vogt-Kölln-Straße 30, 22527 Hamburg, Germany and [3]Universität Hamburg, Center for Data and Computing in Natural Sciences (CDCS), Notkestraße 11, 22607 Hamburg, Germany

Received February 26, 2022; Revised April 05, 2022; Editorial Decision April 10, 2022; Accepted April 19, 2022

ABSTRACT

Upon the ever-increasing number of publicly available experimentally determined and predicted protein and nucleic acid structures, the demand for easy-to-use tools to investigate these structural models is higher than ever before. The Proteins*Plus* web server (https://proteins.plus) comprises a growing collection of molecular modeling tools focusing on protein–ligand interactions. It enables quick access to structural investigations ranging from structure analytics and search methods to molecular docking. It is by now well-established in the community and constantly extended. The server gives easy access not only to experts but also to students and occasional users from the field of life sciences. Here, we describe its recently added new features and tools, beyond them a novel method for on-the-fly molecular docking and a search method for single-residue substitutions in local regions of a protein structure throughout the whole Protein Data Bank. Finally, we provide a glimpse into new avenues for the annotation of AlphaFold structures which are directly accessible via a RESTful service on the Proteins*Plus* web server.

GRAPHICAL ABSTRACT



INTRODUCTION

The Proteins*Plus* (1,2) web server, openly available at https://proteins.plus, offers molecular modeling support for all protein structures that are publicly available as PDB files in the Protein Data Bank (PDB) (3). Usually, workflows for structure-based design necessitate a comprehensive user knowledge of different molecular modeling tools. For example, predicting potential binding sites, finding similar binding sites for ensemble docking, and molecular docking of small molecules of interest into a binding site requires access to and knowledge of a high number of tools with a multitude of parameters. Furthermore, researchers must rely on their computational resources. With the Proteins*Plus* server, these shortcomings are overcome by enabling users to perform all these steps via one unique and easily accessible interface. The server is under constant development including

[*] To whom correspondence should be addressed. Tel: +49 40 428387350; Fax: +49 40 428387352; Email: matthias.rarey@uni-hamburg.de
[†] The authors wish it to be known that, in their opinion, the first six authors should be regarded as Joint First Authors.
Present address: Florian Flachsenberg, BioSolveIT GmbH, An der Ziegelei 79, 53757 St. Augustin, Germany.

fine-tuning, feature extensions, and the integration of additional modeling tools.

Here, we offer insights into feature extensions for the structural multi-purpose comparison tool GeoMine, the newly integrated molecular docking tool JAMDA and MicroMiner - a method that can be used to screen for single-residue substitutions in local protein environments in the whole PDB.

Finally, the artificial intelligence-based protein structure predictions by AlphaFold (currently predicted by AlphaFold Monomer v2.0) enable unprecedented access to high-quality models of proteins of yet unknown structure (4). These models are now readily accessible via the AlphaFold Protein Structure Database (https://alphafold.ebi.ac.uk/) and can be directly imported via the provided REST API.

MATERIAL AND METHODS: EXTENSIONS AND NOVEL TOOLS

GeoMine

From the analysis of binding sites to investigations of geometric preferences for interactions, the ever-increasing number of molecular structures in the PDB offers a multitude of possibilities for in-depth studies of binding sites, their properties and their similarities. This requires comprehensive search capabilities. With GeoMine (5,6), we have developed a search engine that allows for the generation of and the search for atom-based geometric query patterns and an extensive textual and numerical filtering of the PDB. The query atoms can be described manually or automatically with varying degrees of detail, from major properties like the corresponding molecule type, i.e. nucleic acid, protein, ligand, water, or metal, to more restrictive ones, e.g. the molecular surface contribution of a protein or nucleic acid atom. Further feature points like aromatic ring centers can be added to the query and described equally. Distance ranges or hydrogen bond, pi–pi, pi stacking, pi–cation, metal and ionic interactions between atoms and feature points can be introduced into the query, and angle ranges between those can be specified. With the combination of all these features, almost any 3D pattern can be designed and searched in the entire PDB.

In the Proteins*Plus* user interface, the query can be created in a 3D viewer from scratch by the placement of new atoms and feature points or by selecting those in a visualized binding site of a PDB/AlphaFold structure or any uploaded structure file. For this structure, GeoMine predicts binding pockets with interactions and hydrogen atoms using the tools DoGSiteScorer (7) and Protoss (8,9), respectively. If a ligand is present but no pocket has been calculated, a pocket is defined using a radius of 6.5 Å of any ligand atom. The computing times for the iterative search of over one million preprocessed bindings sites depends on the specificity of the query. Most requests can be processed in the range of minutes. For each detected hit, the root-mean-square deviation (RMSD) between the query and the part of the site matching the query is calculated enabling a ranking of the results by geometric fit. The 150 best results are listed in a table and can be visually inspected superimposed to the query in the NGL viewer. Different visualization options are available, for example, choice of residues (complete pocket or only of the residues that match the query). The 150 best-matching pockets can be downloaded in PDB format together with a report containing the statistical overview of all results. The statistics report lists the PDB IDs and ligand names of all found pockets, the distributions of the RMSD values, and the properties of all matched atoms, feature points, distances, interactions, and angles of the query, e.g. the functional group distribution for a matched ligand atom. The user interface with a query history allows a continuous refinement of the results providing an interactive workflow of query modification and subsequent searching in the results. With this tool, protein function or ligand off-targets can be discovered by searching similar binding site properties in 3D space. GeoMine has recently been applied for a detailed analysis of structural features in protein kinase structures (5).

JAMDA

Protein–ligand docking is one of the core tasks in structure-based drug design. With JAMDA, we aimed for the implementation of a fully-automated docking workflow in the Proteins*Plus* server that does not only provide the actual docking algorithm but also encompasses all necessary preprocessing steps, including protonation state assignment and calculation of hydrogen coordinates for the protein (8), prediction of protonation and tautomeric states of the molecules to be docked (10), as well as the generation of 3D coordinates/conformations (11). While a certain degree of manual intervention is possible, our goal was to provide a fully automated workflow with optimized default parameters. This enables even less experienced users to derive potential binding modes of small molecules in the binding site of interest. From the analysis of structure-activity relationships to the test of new binding hypotheses, the established pipeline offers unlimited access to predicted binding modes.

JAMDA docking combines the TrixX docking algorithm (12,13) for initial pose generation with the JAMDA scoring function (14), and our novel LSL-BFGS optimization algorithm(14,15) for scoring and pose optimization. Initially, conformers for the molecule to be docked are generated with the Conformator (11). The raw poses are subjected to a scoring and optimization cascade using the JAMDA scoring function to refine and rank the docking poses.

On Proteins*Plus*, JAMDA allows for a fully automated docking: Only the protein, the binding site, and the molecules to be docked must be provided by the user. The binding site can be defined based on a known ligand or selected from the pocket definitions in Proteins*Plus* (1) (e.g. predicted by DoGSiteScorer (16)). To enable the user to manually adjust the binding sites, all ligand-based and predicted binding sites which do not originate from GeoMine are editable by the user in the pockets tab by clicking on the pencil symbol of the pocket of interest in the upper right corner. Neither the protein nor the molecules to be docked must be manually prepared by the user because this is an integral part of the JAMDA docking workflow: The protein is prepared by assigning likely protonation states using Protoss (8). Furthermore, only structurally relevant water molecules and small molecules that are common cofactors are kept. The molecules to be docked can be provided by picking a ligand from the NGL viewer for redocking studies or by uploading molecules in any common molecular file format (including SMILES without coordinates). Their predominant protonation and tautomeric states are predicted with UNICON (10) prior to docking. Most of these preprocessing steps can optionally be customized by the user.

The preprocessing and docking are performed on the server and, currently, up to five molecules may be docked simultaneously. In the Proteins*Plus* web interface, the resulting docking poses are shown in a table (with JAMDA score and the RMSD if a redocking was performed) and visualized in the NGL viewer panel for interactive analyses. They can also be downloaded for alternative visualizations and further processing. In consequence, JAMDA offers a pipeline for molecular docking that provides reliable results even in the absence of substantial knowledge regarding molecular modeling tools.

MicroMiner

MicroMiner searches for mutations in protein structure databases. On Proteins*Plus*, it screens for single-residue substitutions in the experimental structures of the entire PDB. Retrieved mutant structures can be easily analyzed and compared to the wildtype through automatically generated superpositions in the NGL viewer. The tool focuses on the local 3D micro-environment of single residues in a query protein. It searches the protein structure database for similar local environments with a mutated central residue. For reasonably large wildtype protein structures it is feasible to search for substitutions of all residues in the query at once. In this way, a user can comprehensively explore the wealth of experimental protein structures that exemplify the local effects of mutations through the interactive web interface.

MicroMiner originates from the ASCONA (17) and SIENA (18) technology for binding site similarity search and ensemble compilation. However, instead of focusing on the protein environment of ligands, MicroMiner uses the local 3D micro-environment of any individual residue as the query to search for residues embedded in similar local arrangements. A database search starts by selecting a query residue from which the local 3D protein neighborhood within a distance cutoff (default 6.5 Å) represents the query micro-environment. The connected sequence fragments of this environment are used to identify candidate protein structures with similar sequence fragments in the database. Second, all potential matches are identified by residue-wise sequence alignments. A subsequent fuzzy geometric filter based on the Cα atom orientation and distances of the matching sequence fragments ensures a reasonably similar structural arrangement while tolerating structural changes upon mutation. Thus, we identify local micro-environments with a high sequence and structural similarity. Figure 1 shows the MicroMiner workflow.

Within the Proteins*Plus* server, the user can select single residues of interest or all residues in the input structure to be searched against the PDB. Searching for all residues is feasible within one minute or less on average, depending on the size of the input protein and the number of similar micro-environments in the PDB. The protein structures of retrieved micro-environments can be explored interactively as a structure ensemble in the 3D viewer and sorted by properties of interest, for example, the RMSD of the local environments to investigate the structural effects of mutations. Further applications are the search for highly conserved regions in protein structures, comparisons of the impact of conservative and radical substitutions, or the investigation of structural effects upon substitution for evaluating the reliability and accuracy of computationally generated models of single-residue substitutions.

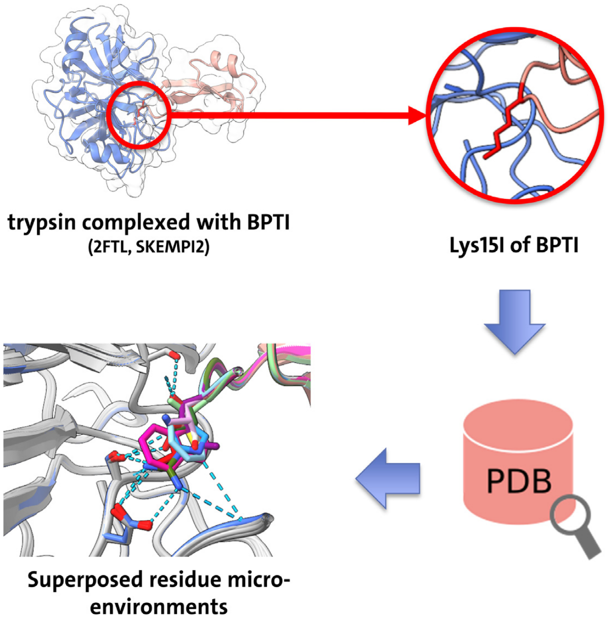

trypsin complexed with BPTI
(2FTL, SKEMPI2)
Lys15I of BPTI
PDB
Superposed residue micro-environments

Figure 1. MicroMiner workflow. With the local 3D micro-environment of a selected query residue, the PDB is searched. Structures from the database containing a similar micro-environment identical in sequence except for the query residue position are retrieved and superposed for analysis. In this way, MicroMiner yields structure ensembles exemplifying the local effects of mutations.

Integration of AlphaFold structures

The inclusion of AlphaFold protein structure models (4) (https://alphafold.ebi.ac.uk/) in the Proteins*Plus* web server enables easy access to machine learning-based predictions of previously unknown structures. The models are accessible on our web server by entering the UniProt Accession Number on the landing page or uploading a preprocessed structure. The user can analyze these structures in the same way publicly available PDB structures can be analyzed by making use of all applicable capabilities of the Proteins*Plus* tools.

Besides the structural uncertainty of AlphaFold structures (19), the missing ligand annotations are a major drawback. This led to the development of the database AlphaFill (20) which annotates the 3D models with cofactors and metal ions and transfers them into the structure assisting in the functional annotation of the models. However, this annotation procedure was only followed for structures that show an identity of at least 35% to known 3D structures

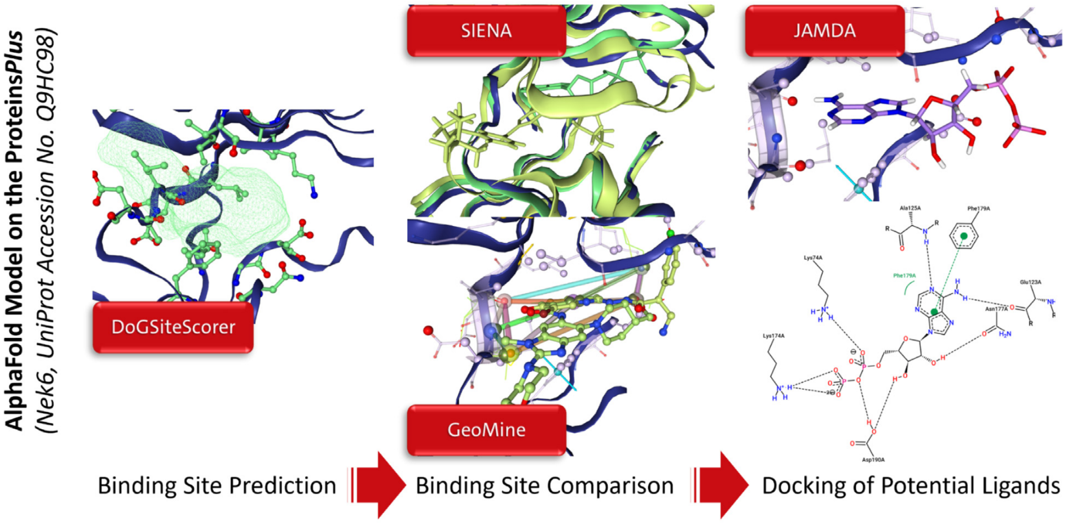


Figure 2. This workflow shows exemplary results for structural investigations of the AlphaFold model for the Nek6 (UniProt Accession Number Q9HC98). First, the user can detect druggable binding sites with DoGSiteScorer. Pocket ‘P_2’ which was predicted as druggable is depicted in green on the right. Next, the pocket can be used for a SIENA search for similar binding sites. Shown are two matches from this analysis with Nek7 structures: 2WQN with ADP and 6S73 in complex with the ligand with the ID F9N in the PDB. GeoMine can be applied for more specific user-defined searches in the binding sites of the PDB. Using a geometric query annotating solvent-exposed potentially interacting atoms and their distances, we found 116 pockets with a similar geometry in the PDB (e.g. cAMP-dependent protein kinase A with the PDB ID 7BAQ, PDB ligand ID T82 or interleukin-1 receptor-associated kinase 4 with the PDB ID 6O94, PDB ligand ID LRS). The corresponding query can be found in the Supplementary Data for upload to the GeoMine tool on the Proteins*Plus* for this structure. Interesting small molecules from the identified similar sites can be downloaded and subsequently be used for molecular docking with JAMDA. The figures on the right show the second highest-scoring predicted binding mode for ADP in the binding site of Nek6 and its 2D interaction visualization with PoseView (21).

stored in the PDB and restricted to common cofactors and ions with potentially functional roles. For researchers interested in the structural annotation of structures that have no known homologs in the PDB, the Proteins*Plus* web service comes in handy. It enables on-the-fly prediction of binding sites with DoGSiteScorer, retrieval of similar binding sites with SIENA, the identification of further potentially interesting ligands by user-defined GeoMine queries, and the molecular docking of these ligands into the AlphaFold model with JAMDA, see Figure 2.

Ligand annotation for AlphaFold models

Given a protein of interest, e.g. the human protein kinase NIMA-related kinase 6 (Nek6), we can start our Proteins*Plus* investigations by providing its UniProt Accession Number Q9HC98 and entering the structural analysis mode of the web service. Next, we can predict potential binding sites using DoGSiteScorer. These predicted sites can be used to search for potential ligands with SIENA. By selecting, for example, the pocket named ‘P_2’ and performing a SIENA search for this predicted binding site, we can retrieve similar sites in complex with various ligands. Besides ADP (the annotation which was also found by AlphaFill), we find similar kinase binding sites in complex with further ligands, in this case, the inhibitor with the PDB ligand ID F9N in complex with Nek2 and Nek7. The active site sequence identity is 94%. The retrieved aligned complexes can be downloaded, together with the corresponding ligand SDF files. The results also enable the exploration of structural flexibility of similar binding sites that can be used, e.g. for the generation of other conformational states that are not covered in the AlphaFold database by homology modeling based on the identified structures.

The ligands retrieved from the SIENA run can either be transferred into the binding site based on the resulting alignment or using the on-the-fly docking tool JAMDA. It can be applied to find whether the found ligands from similar sites can be accommodated in the model’s binding site. However, care should be taken regarding the model quality of the binding site residues as this can have a huge impact on the docking performance. Some preprocessing steps of the original AlphaFold structure might be necessary to obtain reliable ligand binding modes (22).

The search for similar binding sites using the Proteins*Plus*, however, is not restricted to binding sites with a high sequence identity. GeoMine can be applied to generate user-defined queries that search for geometric patterns of interacting binding site residues in nearly one million binding sites (predicted or ligand-annotated) in the PDB. For our example protein kinase, additional GeoMine queries result in the identification of further protein kinases in complex with inhibitors which can be used as idea generators for *in silico* drug design.

SUMMARY AND OUTLOOK

The Proteins*Plus* web server offers a unique access point to protein structure and protein–ligand complex data processing on the worldwide web. Current developments with only conservative extensions of the user interface enable even broader access to molecular modeling tools which usually require comprehensive user knowledge. Furthermore, steady improvements and feature extensions based on suggestions of users render it a lively and well-kept platform. To support users in getting started with the web server, we offer comprehensive documentation of the provided services (https://proteins.plus/help/index) and hands-on tutorials (https://proteins.plus/help/tutorial). As with all computational modeling approaches, the tools behind Proteins*Plus* have their limitations. All users are asked to consult the corresponding methods' publication for more details on the respective restrictions and application domains.

Besides the introduction of new features for GeoMine and the integration of the novel methods JAMDA and MicroMiner, we are in a constant process of elaborating the web server, its tool base, and its potential use cases. The first inclusion of AlphaFold structures in the web server opens new avenues for structural explorations that have not yet been fully explored. With numerous extensions in mind, including 2D and automated query generation in GeoMine or multiple mutations search in MicroMiner, we hope to create a steadily growing, easy-to-use modeling infrastructure for the life science community.

DATA AVAILABILITY

Proteins*Plus* is a publicly available web-based protein structure analysis service, available at https://proteins.plus.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Development of Proteins*Plus* was supported by de.NBI (in part); German Federal Ministry of Education and Research (BMBF) [031L0105 to K.S. and J.S.]; Development of GeoMine was supported by the German Federal Ministry of Education and Research as part of CompLS and de.NBI [031L0172, 031L0105 to K.D. and J.G.]; Development of MicroMiner was supported by the German Federal Ministry of Education and Research (BMBF) as part of protPSI [031B0405B to J.S.]; DASHH: Data Science in Hamburg - Helmholtz Graduate School for the Structure of Matter, funded by the Helmholtz Association [HIDSS-0002 to C.E.]; Center for Data and Computing in Natural Sciences (CDCS), funded by Authority for Science, Research and Equality of the Free and Hanseatic City of Hamburg (BWFGB) [LFF-HHX-03 to A.U.]. Funding for open access charge: Internal university funds.

Conflict of interest statement. Proteins*Plus* uses some methods that are jointly owned and/or licensed to BioSolveIT GmbH, Germany, M.R. is a shareholder of BioSolveIT GmbH.

REFERENCES

1. Schöning-Stierand,K., Diedrich,K., Fährrolfes,R., Flachsenberg,F., Meyder,A., Nittinger,E., Steinegger,R. and Rarey,M. (2020) ProteinsPlus: interactive analysis of protein–ligand binding interfaces. *Nucleic Acids Res.*, **48**, W48–W53.
2. Fährrolfes,R., Bietz,S., Flachsenberg,F., Meyder,A., Nittinger,E., Otto,T., Volkamer,A. and Rarey,M. (2017) Proteins plus: a web portal for structure analysis of macromolecules. *Nucleic Acids Res.*, **45**, W337–W343.
3. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H. and Shindyalov,I.N. (2000) The protein data bank (www.rcsb.org). *Nucleic Acids Res.*, **28**, 235–242.
4. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A. *et al.* (2021) Highly accurate protein structure prediction with alphafold. *Nature*, **596**, 583–589.
5. Graef,J., Ehrt,C., Diedrich,K., Poppinga,M., Ritter,N. and Rarey,M. (2022) Searching geometric patterns in protein binding sites and their application to data mining in protein kinase structures. *J. Med. Chem.*, **65**, 1384–1395.
6. Diedrich,K., Graef,J., Schöning-Stierand,K. and Rarey,M. (2021) GeoMine: interactive pattern mining of protein–ligand interfaces in the protein data bank. *Bioinformatics*, **37**, 424–425.
7. Volkamer,A., Kuhn,D., Rippmann,F. and Rarey,M. (2012) DoGSiteScorer: a web server for automatic binding site prediction, analysis and druggability assessment. *Bioinformatics*, **28**, 2074–2075.
8. Bietz,S., Urbaczek,S., Schulz,B. and Rarey,M. (2014) Protoss: a holistic approach to predict tautomers and protonation states in protein–ligand complexes. *J. Cheminform.*, **6**, 12.
9. Lippert,T. and Rarey,M. (2009) Fast automated placement of polar hydrogen atoms in protein–ligand complexes. *J. Cheminform.*, **1**, 13.
10. Sommer,K., Friedrich,N.-O., Bietz,S., Hilbig,M., Inhester,T. and Rarey,M. (2016) UNICON: a powerful and Easy-to-Use compound library converter. *J. Chem. Inf. Model.*, **56**, 1105–1111.
11. Friedrich,N.-O., Flachsenberg,F., Meyder,A., Sommer,K., Kirchmair,J. and Rarey,M. (2019) Conformator: a novel method for the generation of conformer ensembles. *J. Chem. Inf. Model.*, **59**, 731–742.
12. Schlosser,J. and Rarey,M. (2009) Beyond the virtual screening paradigm: structure-based searching for new lead compounds. *J. Chem. Inf. Model.*, **49**, 800–809.
13. Henzler,A.M., Urbaczek,S., Hilbig,M. and Rarey,M. (2014) An integrated approach to knowledge-driven structure-based virtual screening. *J. Comput. Aided. Mol. Des.*, **28**, 927–939.
14. Flachsenberg,F., Meyder,A., Sommer,K., Penner,P. and Rarey,M. (2020) A consistent scheme for gradient-based optimization of protein–ligand poses. *J. Chem. Inf. Model.*, **60**, 6502–6522.
15. Flachsenberg,F. and Rarey,M. (2021) LSLOpt: an open-source implementation of the step-length controlled LSL-BFGS algorithm. *J. Comput. Chem.*, **42**, 1095–1100.
16. Volkamer,A., Griewel,A., Grombacher,T. and Rarey,M. (2010) Analyzing the topology of active sites: on the prediction of pockets and subpockets. *J. Chem. Inf. Model.*, **50**, 2041–2052.
17. Bietz,S. and Rarey,M. (2015) ASCONA: rapid detection and alignment of protein binding site conformations. *J. Chem. Inf. Model.*, **55**, 1747–1756.
18. Bietz,S. and Rarey,M. (2016) SIENA: efficient compilation of selective protein binding site ensembles. *J. Chem. Inf. Model.*, **56**, 248–259.
19. Perrakis,A. and Sixma,T.K. (2021) AI revolutions in biology. *EMBO Rep.*, **22**, e54046.
20. Hekkelman,M.L., de Vries,I., Joosten,R.P. and Perrakis,A. (2021) AlphaFill: enriching the alphafold models with ligands and co-factors. bioRxiv doi: https://doi.org/10.1101/2021.11.26.470110, 27 November 2021, preprint: not peer reviewed.
21. Stierand,K., Maass,P.C. and Rarey,M. (2006) Molecular complexes at a glance: automated generation of two-dimensional complex diagrams. *Bioinformatics*, **22**, 1710–1716.
22. Skolnick,J., Gao,M., Zhou,H. and Singh,S. (2021) AlphaFold 2: why it works and its implications for understanding the relationships of protein sequence, structure, and function. *J. Chem. Inf. Model.*, **61**, 4827–4831.