Software Tools for Optimization of Biocatalysts

Jan Vilim[a,b], **Zbynek Prokop**[a,c], **and Jiri Damborsky**[a,c], [a] Loschmidt Laboratories, Department of Experimental Biology, RECETOX, Masaryk University, Brno, Czech Republic; [b] Enantis, INBIT, Brno, Czech Republic; and [c] International Clinical Research Center, St. Anne's University Hospital Brno, Brno, Czech Republic

© 2022 Elsevier Ltd. All rights reserved.

Introduction 1
Enzyme Discovery 2
Rational Protein Design 3
Prediction of Protein Structure 3
Engineering Catalytic Activity 4
Targeting Active Site 5
Targeting Access Tunnels 6
Targeting Loops 7
Engineering Thermostability 9
Engineering Solubility 11
Multipurpose Tools 13
Summary and Outlook 15
Acknowledgments 16
References 16

Key Points

- *In silico* tools for the mining and improvement of enzymes are available
- Covered are more than forty relevant and validated software tools
- Focus points from discovery through stabilization to active site re-design
- Substrate access tunnels represent attractive targets for mutagenesis
- Discussed are numerous practical examples related to biocatalysis
- Future opportunities and challenges of the computational enzyme design

Abstract

Computational tools improve protein parameters critical for their utilization in biocatalysis. Here we provide an overview of more than forty software tools available to the community of protein engineers. We discuss the focus points of rational protein design together with the task-specific software options, links to the websites and necessary input files. The success stories from biocatalysis exemplify the application of selected tools to improve practically beneficial enzymes. Finally, we provide a brief overview of unmet challenges and future development in rational protein design.

Introduction

Biocatalysis provides an access to a more sustainable and efficient approach to chemical synthesis.[1] However, the natural proteins suffer from evolution-derived drawbacks since they evolved for catalysis inside the cells living under specific environmental conditions. Our synthetic needs often differ from the requirements of living organisms, we need to improve the properties and efficiency of natural enzymes by protein engineering.[2] Depending on the intended outcome, the improvement of the desired property can be a daunting task. Therefore, computational biochemists and enzymologists require proper tools in the same way as their lab-frequenting colleagues.[3–5]

Over the last thirty years, a vast number of specialized software tools have been introduced to the community and the currently available programs can be used to address the whole workflow of biocatalyst optimization—enzyme discovery from the vast amount of generated genomic data, prediction of the tertiary structure of the target molecule, evaluation of mutations effects on stability, catalytic efficiency, substrate selectivity, and reaction outcome.[3–5] Moreover, some of the tools integrate several types of analysis and provide a simple user-friendly interface, which allows their easy use by the community.

Our book chapter aims at providing a brief overview of software tools that are available to theoreticians as well experimentalists for computational enzyme design. We focus on those tools, which are well established in the research community and the ones that provide integrative computational workflows—combining several approaches in a single environment. Our ambition is not to provide a comprehensive list, thus there will be other alternative tools for specific tasks. Included tools are classified based on their primary use, although many of them are multi-functional. Individual tools are described in terms of their theoretical foundations, functionality and accessibility. Selected case studies are also presented to illustrate some practical examples of what can be achieved by their application to solving specific protein engineering problems. We hope that this book chapter will contribute to the more frequent use of computational tools for enzyme design and discovery.

Enzyme Discovery

Tools able to reliably annotate novel sequences or even predict the catalytic properties of proteins from their sequences constitute an important part of a toolbox of a protein engineer. The discovery process starts with genomic sequencing and subsequent annotation of resulting sequence data. Several remarkable tools have been reported that can successfully predict the reaction type catalyzed by a novel sequence (**Table 1**). Enzymatic reactions are commonly characterized using EC nomenclature.[16] Among the most successful recent contributions, **DEEPre**[6] is a deep learning utilizing tool for the prediction of EC classification. The tool performed well on the first two levels of the characterization, however, increased precision on the third and fourth levels where needed. Improvement was brought in by **ECPred**,[7] which enables EC classification of an enzyme on all four levels of the EC nomenclature using machine learning. This tool was trained and validated using entries from UniProtKB/Swiss-Prot database and provided an improvement in prediction at levels three and four. However, the larger success rate at level four remained elusive, largely due to a limited number of entries necessary for the model training.

To overcome this drawback, the **HEC-net**[8] utilizes the Siamese-Triplet Network approach (STN). The tool was able to significantly outperform the aforementioned approaches at the determination of level 4 of EC number. It achieved 91.2% accuracy and 81.9% F1 score, which is a significant improvement compared to that of 73.8% and 11.3% achieved by ECPred, respectively. It is important to note that enzymes do not have to catalyze only one reaction. The broadening of the capabilities was achieved by **mlHEC-net**,[9] which can add multiple EC numbers if necessary due to the incorporation of a second model.

Bio2Rxn[10] is a web-based software that utilizes several sequence analysis tools to perform a consensus analysis on EC numbers up to four levels using six EC prediction tools. The appeal of Bio2Rxn lies in its high speed (annotation is done in the range of minutes) and a listing of the chemical reaction. This particular feature significantly increases the "friendliness" of this web tool.

Table 1 Selection of tools used for the annotation and discovery of novel proteins.

Tool	*Use*	*Input*	*Link*	*References*
DEEPre	Annotation of enzymatic EC class on two levels	Amino-acid sequence in FASTA format	Webserver http://www.cbrc.kaust.edu.sa/DEEPre/	6
ECPred	Annotation of enzymatic EC class on four levels	Amino-acid sequence in FASTA format	Standalone https://github.com/cansyl/ECPred Webserver https://ecpred.kansil.org/	7
HEC-net mlHECnet	Annotation of enzymatic EC class on four levels	Amino-acid sequence in FASTA format	Webserver http://hecnet.cbrlab.org/	8,9
Bio2Rxn	Consensus-based annotation of EC class on four levels	Amino-acid sequence in FASTA format	Webserver http://design.rxnfinder.org/bio2rxn/	10
EnzymeMiner	Identification of putative members of enzyme families and selection of promising targets for experimental validation	Amino-acid sequence(s) in FASTA format and catalytic residues	Webserver https://loschmidt.chemi.muni.cz/enzymeminer/	11
LIBRA+ LIBRA-WA	Recognition of active and binding sites and corresponding ligands based on the structural information	Protein structure file in PDB format	Standalone http://www.computationalbiology.it/software/ LIBRAv1.zip Webserver http://www.computationalbiology.it/software.html	12,13
GSP4PDB	Search for the published structures that fit the user query of protein-ligand interaction	User-defined graph-based structural pattern	Webserver http://gdblab.com/gsp4pdb/gsp4pdb2/	14
BridgIT	Identification enzymes catalyzing specific reactions	Mol files of ligands and KEGG operators	Webserver https://lcsb-databases.epfl.ch/pathways/Bridgit	15

Another approach to the discovery of novel catalysts is to probe sequences obtained by genome mining for altered properties. This usually stems from previous knowledge about the investigated system. However, it can be challenging to pick a random, previously uninvestigated protein from the database, since solubility-linked expression issues might arise very quickly. For that reason, the software tool **EnzymeMiner**[11] was developed, which enables automated search and annotation of entries in the genomic databases (Fig. 1). EnzymeMiner allows the refinement of search parameters to maximize the coverage of sequence space. This is further visualized in the form of sequence similarity networks and filtration of sequence hits based on the predicted chance of solubility in *Escherichia coli*, identity, presence of transmembrane domains, etc. The validation of EnzymeMiner was performed using the search for novel haloalkane dehalogenase and resulted in the discovery of enzymes with increased activity, thermostability and enantioselectivity.[17]

Another approach that can be used for the discovery of novel biocatalysts lies on the premise that the catalytic site is reaction-specific and its shape can be used as a form of a fingerprint. Tools LIBRA and GSP4PDB take advantage of this. **LIBRA** suite (**LIBRA-WA** server,[12] **LIBRA+** standalone version[13]) is used for active/ligand binding site prediction. In detail, testing of the LIBRA revealed its ability to predict the biologically relevant binding site in 96% of the cases and the correct ligand for the binding site in more than 80% of the cases.

GSP4PDB[14] provides a way to in fact describe the interaction between the ligand and the protein in the form of a graph-based structural pattern (GSP) and its use was validated by the search for PDBs containing the Zinc-finger motif. These tools are quite interesting, however, their use in biocatalysis might be limited for now. However, we assume that with a current spur of machine learning-based methods for structure determination (discussed later), these tools might become very useful for further annotation of catalytic properties of the newly solved structures.

Linked to the description of active sites, **BridgIT**[15] is an interesting tool that utilizes the knowledge of an enzymatic active site to determine the reaction catalyzed by that particular active site. This can be particularly useful, as the authors note, for assigning reactions to the protein sequences in the databases. Particularly interesting is the application of this tool for orphan reactions—reactions that do not have protein assigned, since the method is based on the matching of an active site to the substrate shape. A very nice feature of this tool is its ability to recognize multi-enzymatic reactions. Overall, the discovery of novel biocatalysts is an intrinsic part of biocatalysis.

Rational Protein Design

Prediction of Protein Structure

The use of *in silico* tools for protein engineering is predominantly linked to rational protein design. In general, rational design starts with the careful analysis of the protein properties, particularly the ones linked to its 3D shape, from the viewpoint of the target

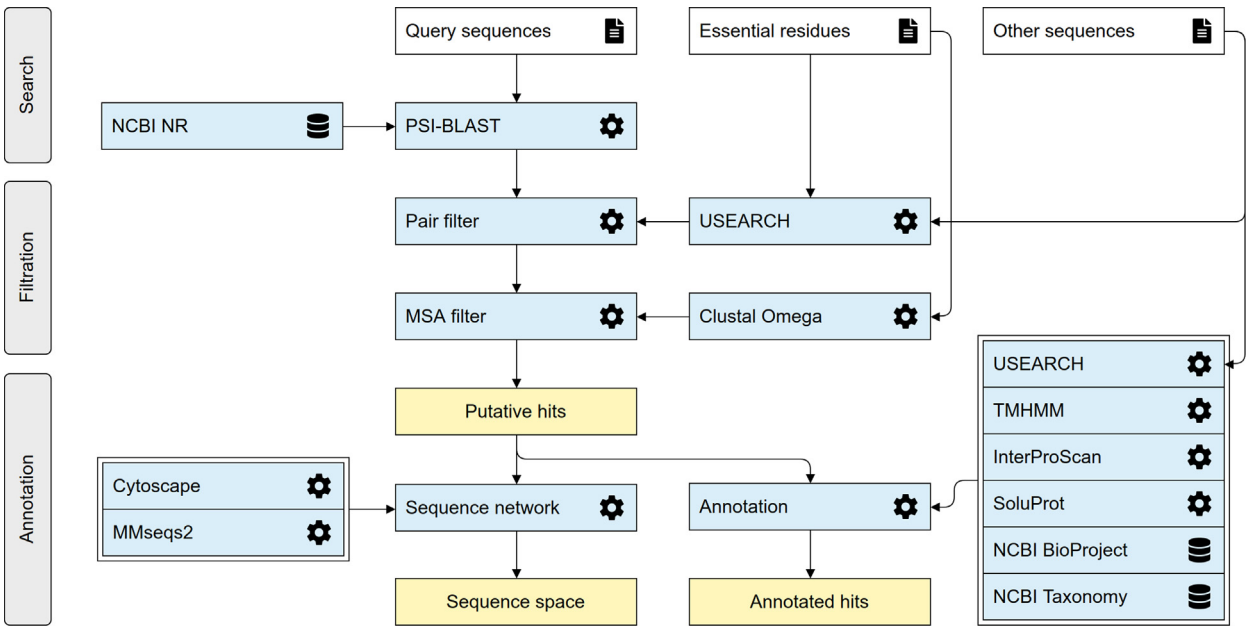


Fig. 1 The workflow diagram of EnzymeMiner 1.0. The workflow consists of three steps: (i) sequence homology search, (ii) filtration of functional sequences, and (iii) annotation of hits. These steps are executed consecutively and automatically. EnzymeMiner requires query sequences and essential residue templates as inputs. The optional inputs are Other sequences that allow EnzymeMiner to calculate the sequence identity between these sequences and all the hits. Input files are highlighted by a white background, tools and databases have a light blue background, and outputs are highlighted by a yellow background. Reproduced with permission from Ref. 11, Oxford University Press.

parameter. However, if no structure is available, this approach is not applicable. Therefore, the structure of the catalyst had to be obtained experimentally, predominantly by the use of X-ray crystallography,[18,19] NMR spectrometry,[20] and cryo-electron microscopy.[21,22]

Experimental determination of protein structures can be a lengthy process and the outcome is not certain. Therefore, over the last thirty years, a great effort has been focused on developing *in silico* approaches that could reliably a rapidly provide highly accurate 3D models of proteins.[23] Regarding the *in silico* approaches, homology modeling has established itself as probably the easiest and the most user-friendly method.[24] It relies on sequence similarity between the protein of interest and previously solved structures available in the PDB database, since proteins that possess high sequence homology in a general fold in the same way. However, the correct prediction of the 3D structure of proteins that bear little or no similarity to previously solved structures has been a long-standing challenge. The prediction tools are compared in a biannual event called Critical Assessment of Protein Structure Prediction (CASP) which is focused on the evaluation of the quality of available tools on a set of modeling tasks. In late 2020, the world has been taken by a storm of news reports that this long-standing problem has been solved by DeepMind during the CASP14.[25] DeepMind's **AlphaFold2** was extremely successful at predicting the 3D structures during the challenge, achieving more than 90% accuracy (Fig. 2). The scientific publication came out in 2021[26] and disclosed that the success of the tool stems from the utilization of machine learning and multiple-sequence alignments that help to uncover the proximity of amino acids due to their mutation rates as well as the physical properties of previously solved structures.

The success of machine learning-based methods was further corroborated by **RoseTTAfold**,[27] a related machine learning-based tool that further built on the principles utilized by AphaFold2. Namely, it uses a three-track neural network that would provide a tighter correlation between the sequence, residue-residue interactions and atomic coordinates. The authors demonstrated the applicability of RoseTTAfold by helping to solve X-ray and cryo-EM structures using molecular replacement.

The aforementioned methods still have space for improvement.[28] However, they are already very valuable for biocatalysis, since quickly generated reliable 3D models of an enzyme provide reach input information for further modeling. It allows for at least preliminary analysis using the tools that are described later in the chapter—the structural information can be further used for the investigation of reaction mechanism, identification of rigid or loose regions of a protein as well as the hydrophobicity or hydrophilicity of the target molecule/region and as such, to streamline the efforts aiming to improve the catalyst at hand without the experimentally verified 3D structure.

Engineering Catalytic Activity

Increasing protein reactivity is arguably one of the main intents of a protein engineer, especially in the field of biocatalysis. The elevation of the performance of the target reaction can be achieved in several ways that are linked either to the chemical reaction in question, such as the modification of the active site and/or its access tunnels, or the modification of the stability of the target biocatalyst, specifically in terms of thermostability, solubility or modification of protein dynamics. A plethora of novel tools of different focus has been developed over the last years. Sadly, their utilization remains relatively scarce and the cross-validation is not always easy. Therefore, we have decided to focus this section on the tools that are either the most prominent or the ones that provide the most benefit in dealing with the often neglected aspects of enzymatic catalysis. Furthermore, the division of this particular section is based on the primary target of the tool to keep it as simple as possible, even though we have to stress that often

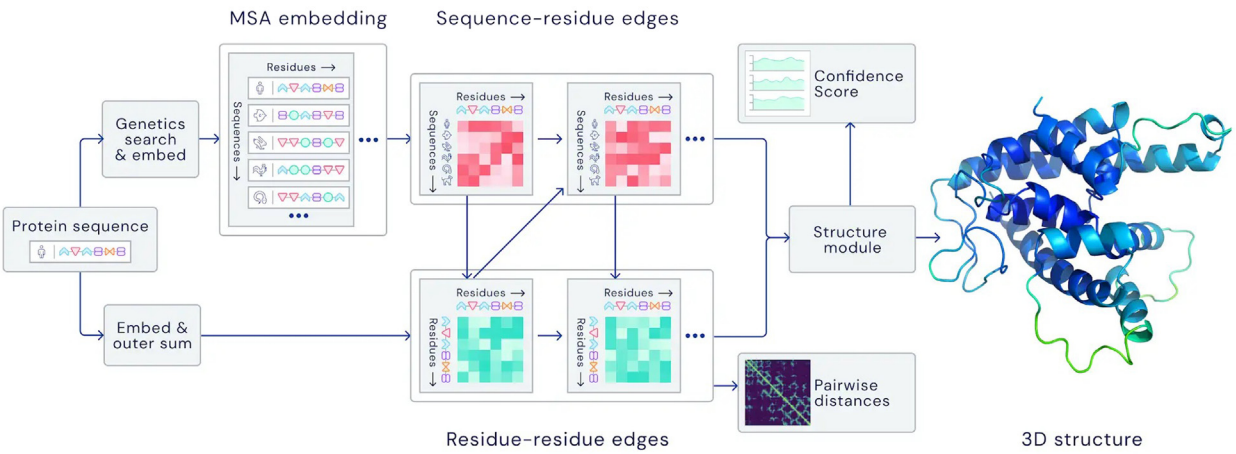


Fig. 2 The workflow diagram of AlphaFold2. The figure illustrates the overall architecture of the network employing multiple sequence alignments and contact maps. The network was trained using the data from publicly available databases and optimized by multiple rounds of calculations employing the supercomputers of Google. Importantly, quantitative confidence scores for the individual predicted regions of the protein structure are calculated. AlphaFold2 is freely available for non-commercial purposes and can be very easily executed using Google Colaboratory (Colab). The figure was downloaded from DeepMind's blogpost (November 2020).

the introduction of mutations influences more than one parameter—e.g. mutations increasing the thermostability can influence the catalytic properties of the enzyme.[29]

Targeting Active Site

The catalytic properties of a biocatalyst are determined by the interplay of the amino acids stabilizing the substrate and its transition state. Therefore, if the catalytic performance of an enzyme is to be improved, the active site is the number one target. As such, visualization and parametrization of the enzyme-substrate binding interactions are of paramount importance. In the last decades, molecular docking has been established as an essential approach allowing researchers to evaluate the compatibility of the active site and the substrate.[30,31] Essentially, molecular docking workflows rely on the evaluation of different binding poses in the target and the calculation of the energetic difference between the binding modes. The resulting values and images provide a road map that either aids researchers to find better binding modes during *in silico* mutagenesis when the wild-type residues are mutated or to rationalize the experimental results.

There are several approaches and a plethora of software tools for molecular docking, and these have been studied elsewhere.[30] Since in this contribution we focus on easy access tools, we will mention two programs that can be used for molecular docking that have obtained general notoriety and that are often incorporated into other tools (**Table 2**). **AutoDock**[32] suite is a well-established software tool enabling various docking-related tasks as well as preparing the input files. Additionally, it incorporates functions for virtual screening, as well as an automatic binding site detection. AutoDock **Vina**[33] is a docking algorithm that is a part of the suite, however, it is an open-source docking algorithm that surpasses the AutoDock, especially from the standpoint of the calculation speed.[45]

Rosie[34] web server is building on the docking capabilities of a Rosetta suite.[46] It incorporates 26 free tools for academic users. To demonstrate its versatility, we would like to mention two recent publications focused on biocatalysis in which the molecular docking/modeling capabilities of Rosie were used. In the first example, Rosie was used to model the protein-protein interactions in

Table 2 Selection of tools used for the improvement of catalytic activity.

Tool	*Use*	*Input*	*Link*	*References*
AutoDock	Molecular docking applications, tools for file preparation and virtual screening workflows	Protein structure file (receptor) and files for target molecules (ligands)	Standalone or incorporated into other tools https://autodocksuite.scripps.edu/	32
AutoDock Vina	Molecular docking applications, are arguably easier to use but lack some functionalities of Autodock suite.	Protein structure file (receptor) and files for target molecules (ligands)	Standalone https://github.com/ccsb-scripps/AutoDock-Vina	33
Rosie	Various molecular docking protocols based on the Rosetta suite	Protein structure file (receptor) and files for target molecules (ligands)	Webserver https://rosie.rosettacommons.org/	34
FuncLib	Improving the catalytic performance of the enzyme employing evolutionary information and Rosetta	Protein structure file, assignment of the active site and catalytic residues	Webserver http://funclib.weizmann.ac.il/bin/steps	35
nAPOLI	Analysis of the protein-ligand interactions	Protein structure file (receptor) and files for target molecules (ligands)	Webserver http://bioinfo.dcc.ufmg.br/napoli/	36
SLITHER	Search for generation of ligand conformations inside the access tunnel	Protein structure file (receptor) and files for target molecules (ligands), docking grid	Webserver http://slither.rcas.sinica.edu.tw/	37
MoMA-LigPath	Analysis of ligand unbinding from the active site	Protein structure file (receptor) and files for target molecules (ligands)	Webserver https://moma.laas.fr/	38
GPathFinder	Analysis of ligand unbinding from the active site	Molecular mol2 files representing both the ligand and the receptor	Standalone https://github.com/insilichem/gpathfinder	39
Caver Web	Identification of protein tunnels and channels and analysis of ligand transport	Protein structure file (receptor) and smiles codes or files for target molecules (ligands)	Webserver https://loschmidt.chemi.muni.cz/caverweb/	40
DeepPocket	Identification of pockets in proteins by deep learning	Protein structure file	Standalone https://github.com/devalab/DeepPocket	41
Caviar	Identification of binding pockets and cavities	Protein structure file in PDB or mmCIF format, trajectory frames from molecular dynamics	Standalone https://jr-marchand.github.io/caviar/	42
DaReUs-Loop	Homology-based insertion of defined loops into the target protein	Protein structure file in PDB format and the amino-acid sequence of the inserted loop	Webserver https://bioserv.rpbs.univ-paris-diderot.fr/services/DaReUS-Loop/	43
LoopGrafter	Transferring one or more loops in between homologous proteins	Protein structure file in PDB format of the scaffold and the insert	Webserver https://loschmidt.chemi.muni.cz/loopgrafter	44

a study focused on the elucidation of biosynthesis of lovastatin.[47] In another example, Rosetta was used for modeling a protein-ligand complex detailing the enantioselective synthesis of α-trifluoromethyl amines catalyzed by P450. Rosetta–based calculations were used for elucidating the rationale behind an enantioselectivity preference of the newly engineered catalyst.[48]

FuncLib[35] is a tool focused on the design of multiple-point mutants that will possess elevated catalytic activity. More specifically, it relies on the evolutionary-sequence analysis and Rosetta atomistic modeling to identify amino acids that are allowed to be present and to model multiple-point mutants that will increase the catalyst efficiency while maintaining stability. The algorithm was tested by the authors on two model cases—phosphotriesterase from *Pseudomonas diminuta* and acetyl coenzyme A synthase from *Salmonella enterica*. In short, the algorithm helped to design several dozens of 3–6 point mutations that possessed 10–4000-fold increased catalytic efficiency compared to respective wild-type enzymes. A very nice example of the power of FuncLib was the engineering of the promiscuous human nicotinamide *N*-methyltransferase into a small set of biocatalysts that can perform selective N-alkylation of pyrazols with excellent region-selectivity. Additionally, some of the variants afforded a second regio-isomer with an excellent selectivity.[49]

nAPOLI[36] is a web server that allows users to investigate interactions between proteins and ligands. However, this tool approaches the problem in a different way than typical docking software, since it mainly serves to discover and describe interactions of the ligand in the binding site and can therefore be used as an interesting starting point for the investigation of the active site. Additionally, due to its ability to process a large amount of data, it can be used to describe also previously unannotated interactions alongside the ones already reported in the literature, as was demonstrated in the case study performed by the authors.

Targeting Access Tunnels

It is important to note, that even if the ligand fits perfectly into the active site, it is not always certain that it will arrive at its destination due to unfavorable interactions somewhere around the access route. Therefore, in the last ten years, there have been published several notable tools that are focused on the identification of access tunnels that play role in the transport of substrate to the active sites.[3] This is a very relevant problem for biocatalysis, therefore, we have devised a small selection of available tools (Table 2).

Slither[37] is an interesting tool that utilizes the power of molecular docking (enabled by AutoDock[32] and MEDock[50]), to calculate the binding energy between the ligand and the protein across the access tunnel of the active site. This can be particularly interesting for the identification of the bottlenecks in access pathways for a new substrate or retrospective rationalization of the effect of mutations localized outside the active site.

MoMA-LigPath[38] utilizes a geometrical approach based on the Manhattan-like Rapidly-exploring Random Tree (ML-RRT), which treats both ligand and the residues as partially flexible and tries to find the path by which the ligand leaves the active site. The generated output contains a visual representation of the possible unbinding pathways. However, it does not contain any information about energy changes during the unbinding.

GPathFinder[39] is an interesting tool belonging to the GaudiMM suite and as a part of it, it utilizes the aforementioned AutoDock. Compared to the tools mentioned previously, it allows finding both binding and unbinding pathways of the ligand. The authors studied its abilities in three different case systems—transport of a molecule of glycerol through an aquaporin, unbinding of a suicide inhibitor from human indoleamine-2,3-dioxygenase and search for access channels if a human cytochrome P450 2C19. The algorithm was able to find previously described access tunnels and channels as well as some new ones. Additionally, its earlier version was used to uncover the catalytic mechanism of an exo-hydrolase from a *Hordeum*.[51]

Caver[40,52,53] suite comprises three tools that allow the identification of channels, tunnels as well as binding trajectories and energies. Here we will particularly focus on the **Caver Web**[40] since it builds on top of the CaverDock[53] and Caver[52] itself. This web-based tool offers a way to detect channels and tunnels as well as a catalytic pocket using Caver, and subsequently, perform an optional ligand transport analysis using CaverDock software (Fig. 3). The authors used CaverWeb in different study cases. For example, the analysis of the entrance tunnels of different haloalkane dehalogenases showed a correlation between the size of the active site tunnel and the preferred size of the accepted substrate. Another example of a possible use reported by the authors was a virtual screening of leukotriene A4 hydrolase/aminopeptidase inhibitors since CaverWeb allows docking alongside the residues along a tunnel, which in turn significantly enriched the sampling region. The virtual screening uncovered a possible tunnel-dependent inhibition mechanism linked to some of the tested molecules, which would not be possible with the classical approach targeting solely the active site. The tool stands to be quite useful indeed. There are two publications not originating from the group of authors that report the successful utilization of CaverWeb toward the improvement of the catalytic activity of a protein. More specifically, it was used to identify water access tunnels in the study that was focused on increasing of the coupling efficiency of P450 BM3 from *Bacillus megaterium*, in which the change of the tunnel polarity resulted in the lesser presence of water molecules in the proximity of heme in its active site which correlated with the increased coupling efficiency.[54] The tool was further utilized to detect tunnels in a study focused on the engineering of the acyltransferase from *Mycobacterium smegmatis*.[55] The detailed Caver Web workflow can be found in Fig. 3.

DeepPocket[41] is an interesting tool focused on the detection of binding sites. Specifically, it utilizes a 3D convolutional neural network. In detail, the software strips the input structure of any heteroatoms and solvent molecules and searches for pockets using the FPocket[56] (which is also present in the Caver Web). The resulting list of pocket centers is subsequently processed and ranked using the convolutional neural network. The performance was tested using four datasets. It appears that the use of CNN together with the FPocket enabled a more precise determination of the pockets in the evaluation datasets compared to other tested methods, therefore, this particular tool might serve as an addition to the FPocket in bioinformatics workflows.

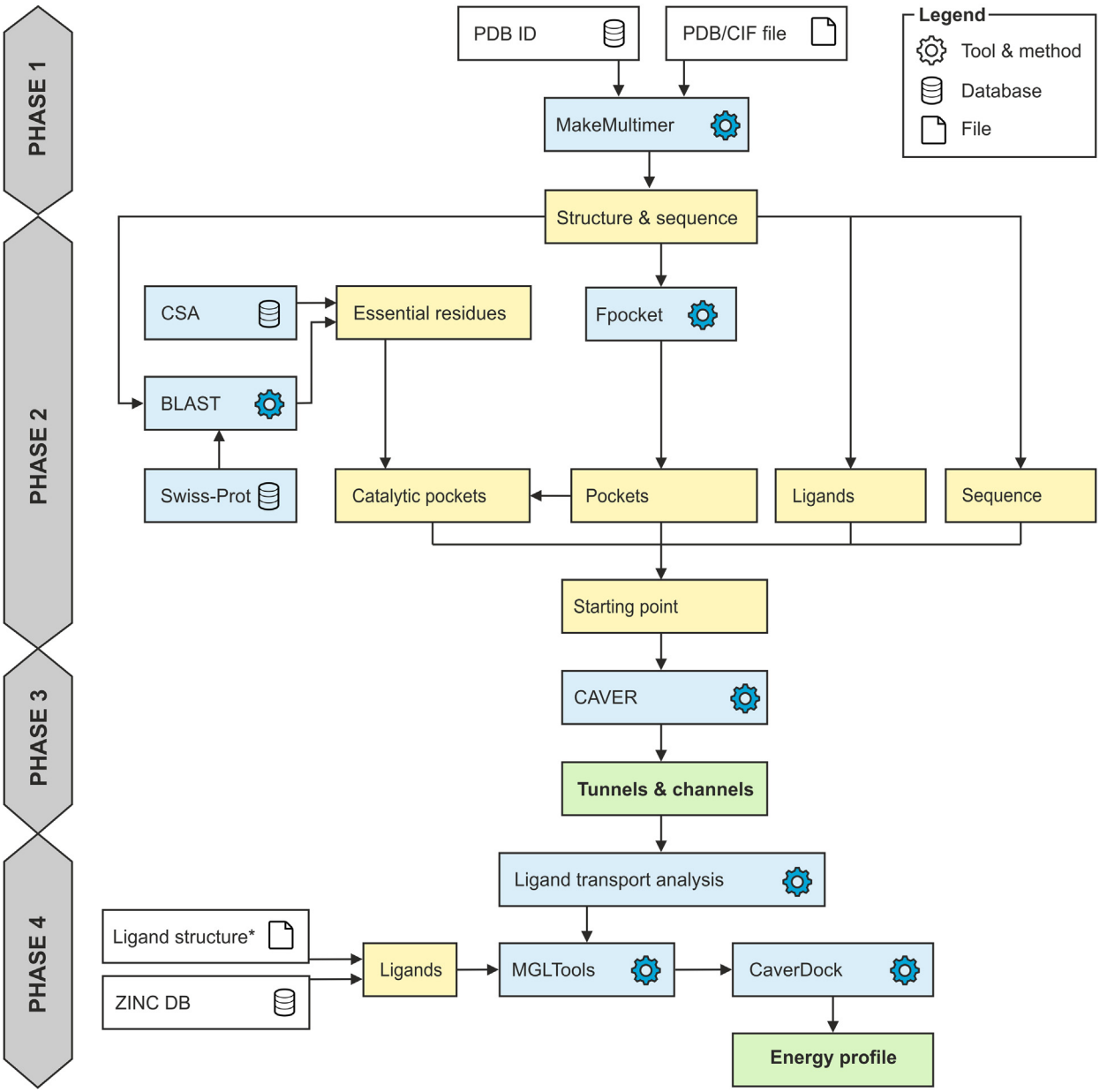


Fig. 3 Workflow diagram of Caver Web 1.0. The process consists of four phases: (1) structure pre-treatment, (2) detection of the starting point, (3) identification of tunnels/channels and (4) analysis of ligand transport. Ligand structures can be uploaded in all formats supported by OpenBabel. Reproduced with permission from Ref. 40, Oxford University Press.

Caviar[42] is another recent addition to the active site/cavity detection tools. It is a standalone software that is intended for more Python-familiar users and as such, it can be more useful for tool developers, similar to the previously mentioned deep pocket. Compared to the tools dealing with the ligand-pathway search, the last two tools might not be very useful directly, however, they can provide better basal cavity detection, which can in turn provide an increased performance of the path-finding tools. Therefore, we are quite curious about the improvements to the methods that we will see in the future.

Targeting Loops

Loop engineering is another interesting approach that can lead to improved performance of the biocatalyst. Overall, this particular kind of enzyme engineering has been known for quite some time as it can be considered one of the key tools in nature's repertoire that leads to the evolution of novel proteins.[57] This fact is well known and it has been already utilized for the redesign of enzymes. The introduction of loops resulted in a rapid activity switch in glyoxalase II, which was transformed into β-lactamase.[58] In terms of biocatalysis, a very nice example of loop remodeling was reported by the Arnold group. In this particular contribution, molecular

dynamics simulation was used to discover the flexibility of a particular loop in Cytochrome P450 TexT.[59] This information ultimately led to the development of a variant that possessed altered regioselectivity of the target reaction.

DaReUs-Loop[43] is a web-based tool that was developed from its initial standalone version[60] to provide easy access to loop remodeling in the homology models. The webserver utilizes PDB file and a sequence as an input and offers three types of analysis—remodeling, modeling and advanced. Effectively, it can model up to 20 loops in one template that can be between 2 and 30 amino acids in length. There are few input options, but the server is incredibly user-friendly and the outcomes are quite easy to understand. However, it is solely focused on the homology model application, therefore, it cannot consider the dynamic parameters of the loop. Additionally, it does not provide scoring of the final designs.

LoopGrafter[44] is a recently published tool that deals with the modeling of loops and is intended to be used for the improvement of the performance of the target enzyme. The server utilizes two input PDB files—the scaffold protein which is going to receive a new loop and the insert which is going to provide the loop. The web tool navigates the user through six steps (Fig. 4)—secondary structure definition, loop exploration, flexibility evaluation, correlation evaluation, loop pairing and finally, loop grafting. The graphical interface of the tool seamlessly navigates the user to the final calculation with up to 1000 designs listed and scored using Modeller's DOPE or Rosetta's FastRelax scores. The authors applied their loop-grafting strategy to transplant a loop from *Renilla* luciferase to a structurally related bifunctional ancestral luciferase. The resulting chimera possessed 40,000-fold elevated fluorescence activity compared to the scaffold enzyme. Additionally, the grafting process also resulted in higher substrate affinity as well as lesser product inhibition. The comparison with the aforementioned DaReUs-Loop revealed that LoopGrafter can create novel chimeric sequences, whereas the DaReUs-Loop requires the full sequence of the final chimera to be defined in the input (Table 2).

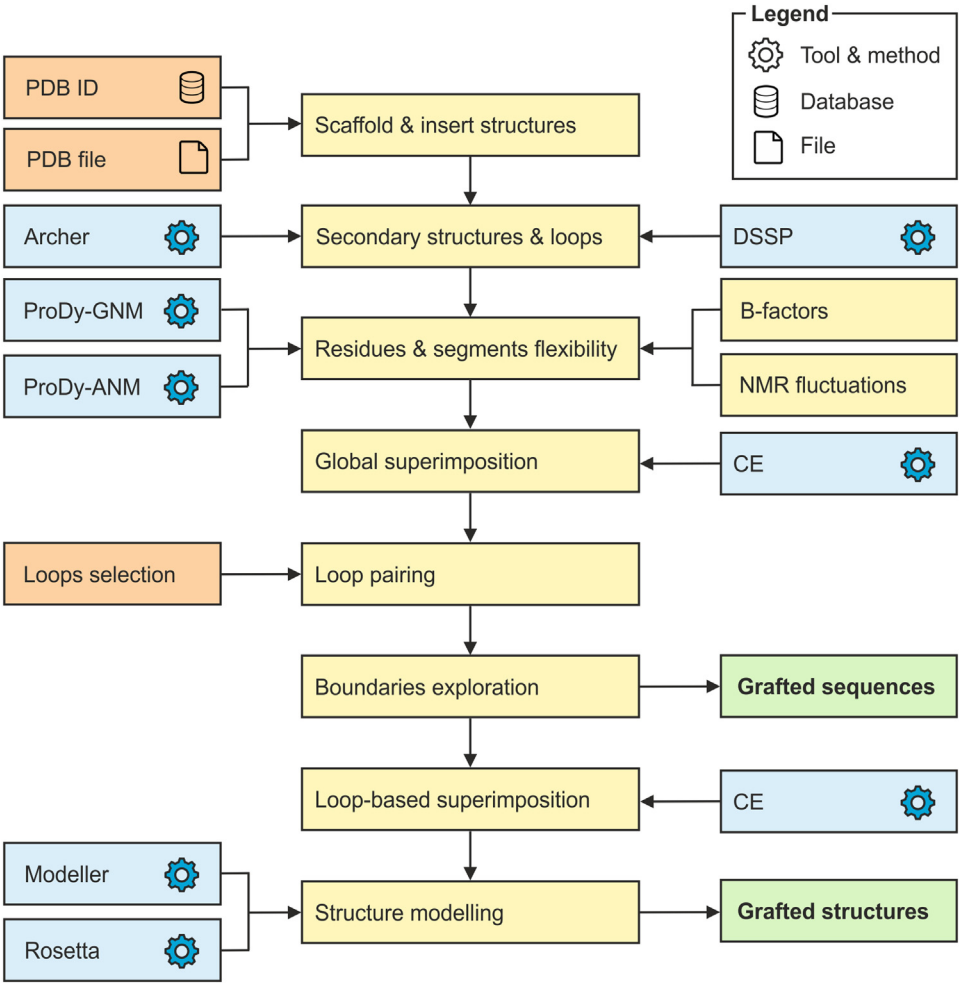


Fig. 4 Workflow diagram of LoopGrafter 1.0. Scaffold and insert tertiary structures can be fetched from the PDB database or uploaded by the user. DSSP and Archer definitions are used for the calculation of secondary structures and loop geometries. Elastic Network Models implemented in ProDy-GNM and ProDy-ANM provide flexibility and cross-correlations. Flexibility analyses are followed by structural superimposition using CE and pairing of loops in the two input proteins. A list of designed sequences is generated after a systematic exploration of possible grafting boundaries. 3D models for each of the designed sequences are generated and evaluated with Modeller's DOPE and Rosetta's FastRelax scores. Reproduced with permission from Ref. 44, Oxford University Press.

Engineering Thermostability

The increased thermostability of biocatalysts often translates to a longer life span in the reaction mixture, which should transform into a greater yield of the reaction. Together with the fact that the size of a stabilizing or destabilizing effect of a mutation is virtually impossible to guess by "visual analysis", software tools enabling the increase of thermostability of any protein are in high demand. In general, the various tools reported in the literature rely either on machine learning to unearth hidden dependencies or force-field dependent energy calculations or evolution-based approaches leading to more stable variants. The tools mentioned in this section are summarized in **Table 3**.

FoldX[61] has been the benchmark to which the tools focused on stabilization are often compared. FoldX is an empirical force field that serves to evaluate the effects of mutations on total free energy. Over the years, it gave rise to the whole suite of tools that are ready at the disposal of protein engineers. Its importance can be further stressed by the sheer amount of citations that the original publication amassed since its release. Currently (May 2022), there are 1400 citation records in Scopus. An example of FoldX being used as a standalone tool can be the engineering of β-aminotransferase by Rabe and colleagues.[61] In their contribution, the FoldX was used to design a two-point mutant with a melting temperature increased by 4 °C and a three-fold increased half-life.

FRESCO[62] is a workflow that helps to design stable variants using FoldX and Rosetta. It was used to design a more stable variant of cyclohexanol monooxygenase from *Rhodococcus* sp. HI-31.[72] The resulting variant bearing eight mutations exhibited a melting temperature increased by 13 °C together with a 33-fold increase of half-life at 30 °C. In another example, FRESCO was used to design a robust ω-transaminase.[73] In that particular contribution, the best 6-point variant exhibited an increased melting temperature by 25 °C. Additionally, it was also more stable in presence of isopropanol as a co-solvent and exhibited increased catalytic activity.

PROSS[63] enables the quick and user-friendly design of proteins possessing elevated stability. In the original publication, the authors demonstrate the use of the algorithm by the increasing the thermostability of an acetylcholine esterase from humans (hAChE), therefore, the tool is listed in the section focused on the thermostability, however, we have to note that the focus of PROSS is larger and can be used to also improve expression and the solubility of the protein as well as their activity. In detail, the tool utilizes mutational scanning, in which the amino acids that are allowed to a certain position are derived from the amino acids that can be observed in the homologous sequences and are subsequently evaluated using Rosetta. This particular approach allowed authors to avoid false positives and to increase the expression of hAChE variant by almost 2000-fold compared to the wild-type. The

Table 3 Selection of tools used for the improvement of protein thermostability.

Tool	*Description*	*Input*	*Link*	*References*
FoldX	Force-field describing the effect of mutations on the stability of proteins	Protein structure file in PDB format	Standalone https://foldxsuite.crg.eu/	61
FRESCO	Stabilization of proteins by the design of mutant libraries	Protein structure file in PDB format	Standalone https://groups.google.com/g/fresco-stabilization-of-proteins	62
PROSS	Design of more soluble and stable proteins focused on the heterologous expression in bacteria	Protein structure file in PDB format	Webserver https://pross.weizmann.ac.il/step/pross-terms/	63
ProTSPoM	Identification of destabilizing amino acids in the protein structure	Protein structure file in PDB format	Webserver http://cosmos.iitkgp.ac.in/ProTSPoM/	64
FireProt	Design of more stable protein variants using energy- or evolution-based analysis	Protein structure file in PDB format	Webserver https://loschmidt.chemi.muni.cz/fireprotweb/	65
DynaMut	Analysis of a structure and further evaluation of the putative influence of mutations on the protein dynamics	Protein structure file in PDB format	Webserver http://biosig.unimelb.edu.au/dynamut/	66
TKSA-MC	Design of more stable variants by the optimization of protein charge-charge interactions	Protein structure file in PDB format	Webserver http://tksamc.df.ibilce.unesp.br/	67
pStab	Evaluation of the influence of single- and multiple-point mutants on protein stability parameters	Protein structure file in PDB format	Webserver https://pbl.biotech.iitm.ac.in/pStab/	68
ThermoNet	Evaluating the influence of single-point mutations on the protein stability	Protein structure file in PDB format	Standalone https://github.com/gersteinlab/ThermoNet	69
Yosshi	Stabilization of proteins by the introduction of disulfide bridges	Protein structure file in PDB format and the MSA of a protein family	Webserver https://biokinet.belozersky.msu.ru/yosshi	70
FireProtASR	Design of stable mutants using ancestral sequence reconstruction approach	Single protein sequence or multiple sequences or multiple sequence alignment	Webserver https://loschmidt.chemi.muni.cz/fireprotasr/	71

usefulness of PROSS was further demonstrated with peroxidases. The resulting variants were successfully expressed in yeast whereas the parent enzymes were not. The variants possessed elevated stability as well as functional diversity.[74] PROSS was further validated using community-wide evaluation.[75] PROSS enhanced expression and solubility, thermostability and activity of 14, 10 and 7 target proteins, respectively. PROSS contributed to an increase in both soluble expressions (64% in low-throughput screening). Ultimately, elevated thermostability was observed in 9 out of 10 cases.

ProTSPoM[64] utilizes FoldX force field to predict the effect of single point mutations on the thermodynamic stability of a protein. However, it can be postulated, that the success rate of free-energy calculation-based methods significantly drops when the calculation contains multiple point mutations at the same time.

FireProt[65] is a web-based computational workflow that combines sixteen individual tools and three protein strategies for the fully automated design of stable proteins (Fig. 5). In detail, it divides the workflow into three phases—annotation, prediction of single point mutations and design of a thermostable protein and it delivers a validated multiple point mutant designed by energy-based workflow, the evolution-based workflow or a combinatorial mutant. To minimize the incorporation of mutation pairs having an antagonistic effect, the protocol utilizes the incorporated Rosetta tool to avoid possible clashes. FireProt has been experimentally validated on three proteins – haloalkane dehalogenase DhaA, γ-hexachlorocyclohexane dehydrochlorinase LinA and fibroblast-growth factor 2. Remarkably, the tool provided stabilization of 25, 21 and 15 °C, respectively. As a further demonstration

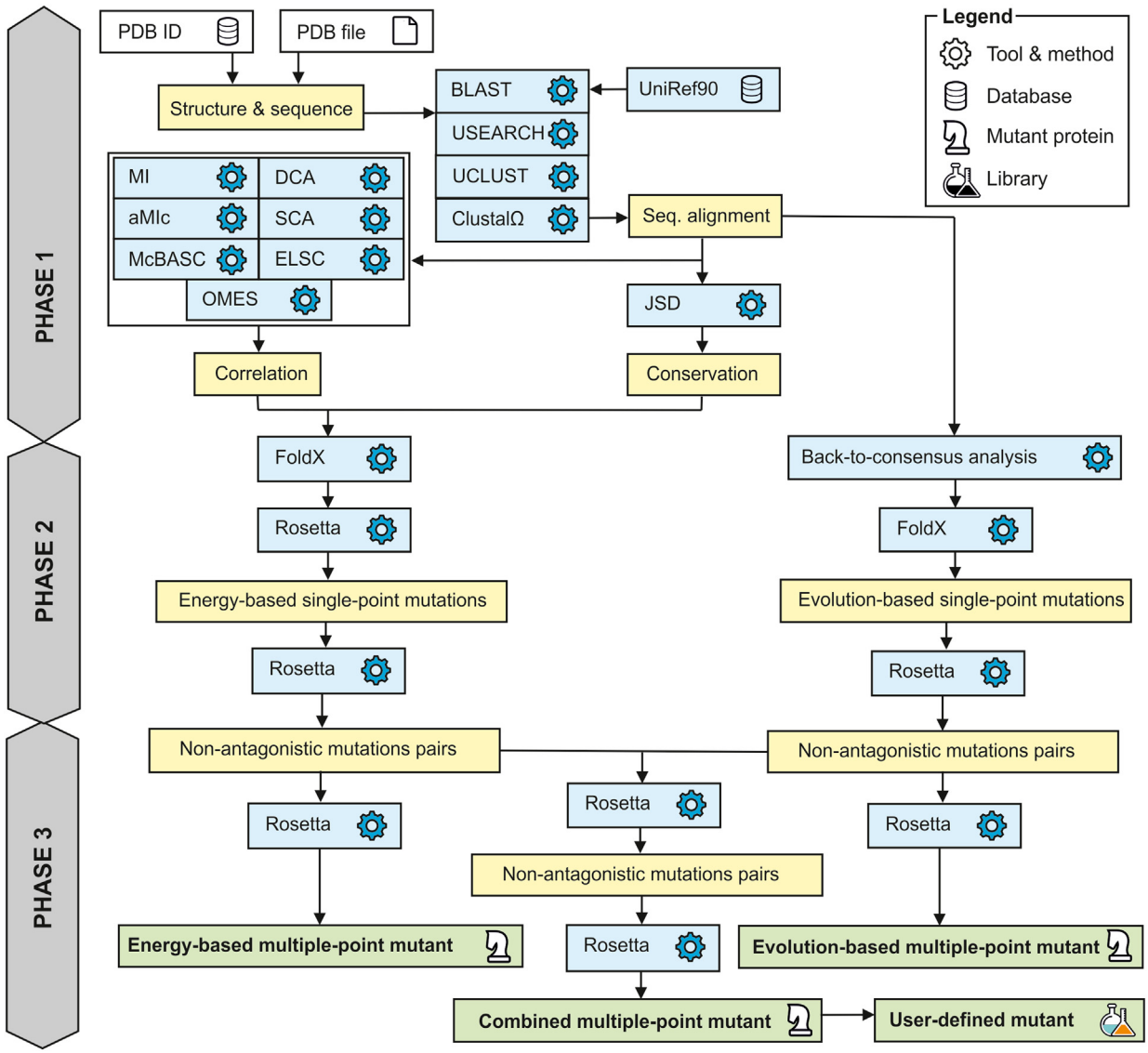


Fig. 5 Workflow diagram of FireProt 1.0. A protein defined by the user is initially annotated using several prediction tools and databases (Phase 1). The energy-based and evolution-based approach is applied to assemble a list of potentially stabilizing single-point mutations (Phase 2). Finally, three multiple-point mutants are generated in an additive manner, while removing potentially antagonistic effects of mutations (Phase 3). FireProt 2.0 is currently in beta testing. Reproduced with permission from Ref. 65, Oxford University Press.

of its usability, the tool was used to stabilize a ketoreductase ChKRED12.[76] The resulting quadruple mutant exhibited increased temperature optimum by 5 °C and more than 21-fold increased half-life compared to the wild type. Significantly enhanced FireProt 2.0 is currently in beta testing. This new version provides low-risk and high-risk design, several new modules for stabilization (ancestral sequence reconstruction, B-factor analysis, saturation mutagenesis, suppression of antagonistic effects, etc.), AlphaFold2 and ProMod3 models, visualization in Mol*, and many other improvements.

DynaMut[66] is another interesting web-based tool that serves to design stable proteins. Specifically, this tool allows either analysis of protein dynamics or effects of target mutations on dynamics and stability, using a consensus of three computational methods which leads to increased performance when compared to the stand-alone tools, sadly, its utilization in biocatalysis remained scarce.

TKSA-MC[67] represents a different approach to protein stabilization. This particular web server serves for the optimization of charge-charge interactions. This is done by the evaluation of protein surface accessibility and its protonation states. The outcome is presented in a graph form and provides an insight into mutations that contribute to the increase of the free energy and as such, their mutation to a different amino acid should result in greater stability. The tool was verified experimentally, and was quite successful, especially in the case of a heat-shock protein from *Bacillus subtilis*, where a single mutation raised the melting temperature by 8 °C.

PStab[68] is another approach utilizing the calculation of charge-charge interactions. The web server serves for the identification of possible stabilizing mutations under a variety of different conditions (temperature, pH, ionic strength, for example) and subsequently performs virtual mutagenesis up to quadruple mutants. Furthermore, it can generate unfolding curves and it can be used to identify mutational hot spots as well as to generate fitness landscapes.

ThermoNet[69] is a deep 3D convolutional neural network able to predict the mutation-induced changes of ΔΔG. Treats the 3D structure of a protein as a multichannel voxel grid and uses pre-defined biophysical rules (hydrophobicity, aromaticity, hydrogen bonding, ionizability and occupancy) to further parametrize each voxel and these are further used as an input to the trained neural network. Interestingly, ThermoNet was able to outperform commonly used tools in the determination of ΔΔG upon the point mutation.

Yosshi[70] is a web tool that enables locating of the most promising positions for the incorporation of disulfide bridges based on the input PDB and a multiple sequence alignment of a set of proteins with diverse functions but a related structure and low sequence identity. There are no biocatalysis-related applications, but it can be an interesting approach or a quick stabilization of proteins with a low melting point. Related to the multiple sequence alignment, the ancestral sequence reconstruction is used for the identification of protein sequences that should possess higher stability, based on the idea that earlier versions of the proteins had to work under harsher conditions. There are several available tools, however, they tend to be rather hard to use for an inexperienced user.

FireProtASR[71] offers an automated workflow with an easy to use web-interface that allows to the construction of ancestral sequences with only an input sequence (**Fig. 6**). The first phase utilizes the aforementioned EnzymeMiner to collect relevant homologous sequences and several other tools to then construct a phylogenetic tree containing a maximum of possible genetic diversity. In the second step, the ancestral sequences are generated based on the users' input regarding the evolutionary matrix, organism and the nodes of the phylogenetic tree. The workflow has been validated on haloalkane dehalogenase DhaA. Out of the six ancestral variants, five were expressed in soluble form in *E. coli* and all of them possessed enhanced thermostability by more than 20 °C. Additionally, all five mutants retained the dehalogenase activity and in three cases, the activity was increased. Remarkably, three of these ancestral proteins exhibited an additional monooxygenase activity. Overall, the above-mentioned methods represent the state of the art of *in silico* tools that enable the stabilization of target proteins. The biocatalysts exhibiting enhanced stability can be used for the intensification of target reactions and they can significantly contribute to the widespread utilization of biocatalysts in chemical processes. Since there are tools that are easy to use also for non-experts (computational chemists or evolutionary biologists) we assume they will gain more popularity in the coming years.

Engineering Solubility

The solubility of a biocatalyst relates to its ability to remain solvated in a reaction buffer. In a broader sense, it also relates to the heterologous expression of correctly folded and active protein.[77] It is very important to have a soluble protein for the development of efficient biocatalytic processes. Several tools have been developed to predict or improve the solubility of an enzyme from its sequence (**Table 4**).

ESPRESSO[78] is one of the benchmark tools for protein solubility prediction. It utilizes two different approaches, the sequence property-based method and a sequence pattern-based method. The first approach utilizes 437 features that are extracted from the nucleotide and amino acid sequence and the second determines sequence patterns associated with expression and solubility. The useful features lie in both the ability to estimate the feasibility of successful heterologous expression in *E. coli* or wheat germ cells but also to suggest a target region for the expression/solubility improvement.

DSResSol[79] is a deep learning-based solubility prediction tool that combines squeeze excitation residual networks with dilated convolutional neural networks. It utilizes 85 features that are extracted from the sequence such as hydrophobicity, secondary structure, molecular weight and charge, among others. The combination, according to the authors, enables the extraction of difficult-to-note long-distance interactions. On the available test set, it outperformed the other solubility predictors.

DeepSol[80] is a tool based on deep learning that predicts the solubility of target proteins from a sequence. This deep-learning approach uses a convolution neural network trained on 28,972 soluble and 40,448 insoluble proteins. Ultimately, the tool was

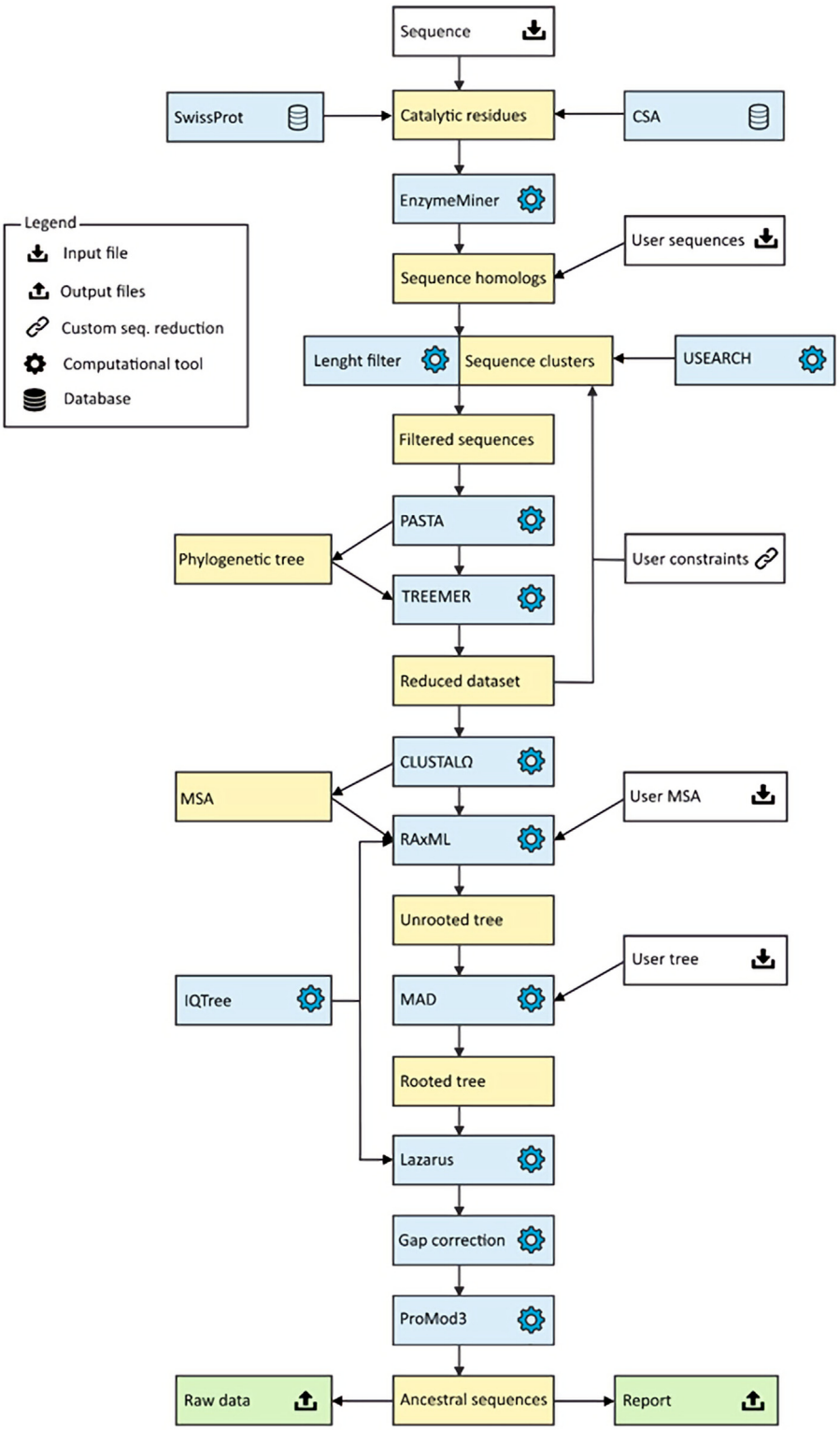


Fig. 6 Workflow diagram of FireProtASR 1.0. To infer ancestral sequences in a fully automated way, a set of biologically relevant homologous sequences are collected from genomic databases and reduced to a suitable size (Phase 1). Next, several state-of-the-art methods are utilized to construct a multiple sequence alignment and a phylogenetic tree, which are then used to support the inference of ancestral nodes and reconstruction of ancestral gaps (Phase 2). The FireProtASR requires no user intervention beyond providing a query sequence and selecting catalytic residues used to identify a biologically relevant set of homologous sequences. However, it is also possible to start a calculation with a user-defined initial set of homologous sequences, MSA, or a phylogenetic tree instead of a single sequence. Reproduced with permission from Ref. 71, Oxford University Press.

Table 4 Selection of tools used for the improvement of protein solubility

Tool	*Use*	*Input*	*Link*	*References*
ESPRESSO	Prediction of protein expression and solubility based on the secondary structure features and sequence patterns	Amino-acid sequence in FASTA format	Webserver http://mbs.cbrc.jp/ESPRESSO	78
DSResSol	Prediction of protein solubility using sequence patterns	Amino-acid sequence in FASTA format	Standalone https://github.com/mahan-fcb/DSResSol	79
DeepSol	Deep-learning prediction of the expression and the solubility based on the secondary structure features or the sequence patterns	Amino-acid sequence in FASTA format	Standalone https://zenodo.org/record/1162886#.Yn5WOehByHs	80
SoluProt	Prediction of the expression and solubility of target protein in *E. coli* based on machine learning	Amino-acid sequence in FASTA format	Webserver https://loschmidt.chemi.muni.cz/soluprot/	81
SOLart	Prediction of solubility using structure-related features	Protein structure file in PDB format	Webserver http://babylone.ulb.ac.be/SOLART/k_query.php	82
SoDoPe	Prediction of the expression and the solubility of a protein in *E. coli*, together with suggestions for improvement	Nucleotide or amino acid sequence in FASTA format	Webserver https://tisigner.com/sodope	83
AGGRESCAN3D	Prediction of protein solubility and *in silico* mutagenesis for suppressed aggregation	Protein structure file in PDB format	Webserver http://biocomp.chem.uw.edu.pl/A3D2/	84

able to match the performance of the state-of-the-art prediction methods and even surpass them after incorporating knowledge about the biological features linked to the particular sequences.

SoluProt[81] is a sequence-based tool that enables the prediction of solubility in *E. coli*. For the prediction itself, it uses 250 parameters that are linked to the features in the sequence, such as amino acid composition, secondary structure content, flexibility, physicochemical attributes and several others. On the test set of 3100 proteins, the tool outperformed other state-of-the-art methods, including the aforementioned ESPRESSO and DeepSol. SoluProt has also been integrated as a part of an EnzymeMiner workflow where it helps to prioritize sequences that have the highest chance for successful expression in *E. coli*.

SOLart[82] relies predominantly on structural information, which makes it different from the aforementioned methods utilizing the information contained in the sequence. The tool utilizes 103 structure-based features together with 28 purely sequence-based features. The authors decided to validate SOLart on two independent tests test of the 3D structure of *Saccharomyces* proteins.

SoDoPe[83] uses protein structure for solubility prediction. Interestingly, it utilizes the solubility-weighted B-factors for its calculations. This approach stems from the assumption that flexibility and stability are inherently connected[85] and the B-factor is a reliable predictor of flexibility in protein structures. During the validation phase, the tool achieved reasonable performance, which is remarkable considering that it operates with a single parameter.

AGGRESCAN3D 2.0[84] is a web server that focuses on the prediction and engineering of protein aggregation and solubility. Aggregation is a medically relevant problem, however, it can still be a very useful parameter when trying to optimize protein expression yields. It helps to identify single point-mutations that have an anti-aggregation effect on the target structure. It utilizes FoldX to introduce mutations and then it provides a calculated output in a form of an aggregation score. Aggregation-prone regions are visualized in form of a color-coded 3D structure.

While these tools are getting more sophisticated and precise, there are not a lot of direct biocatalysis-related reports of their use. While desirable, the biocatalysis community does not seem inclined to utilize these tools. The tools are not able to deliver proteins that are easy to express and soluble in organic solvents. The solubility in organic solvents remains a large hurdle preventing wider utilization.[86] Many solubility-focused tools outperform their opponents, however, it appears that these comparisons are not very reliable. Proper experimental validation of the tools by their wider usage is therefore necessary.

Multipurpose Tools

Multipurpose tools allow users to perform different tasks as a part of one tool or provide multiple tools as a part of one suite developed in one organization (Table 5).

Rosetta is probably the best example of a comprehensive suite covering diverse aspects of computational biochemistry. In this contribution, Rosetta was already mentioned in the section that detailed the improvement of catalytic activity by various methods. Rosetta offers a large selection of tools that can be used for protein modeling with a focus on improving the activity and stability, as well as *de novo* design of new enzymes, which however requires multiple rounds of directed evolution, for achieving practically useful catalysts. In 2020, the capabilities of the Rosetta suite have been comprehensively summarized in Nature Methods.[46]

Table 5 Selection of multi-purpose tools.

Tool	*Functionalities*	*Link*	*References*
Rosetta suite	Molecular docking, molecular modeling, protein design	Standalone https://www.rosettacommons.org/home	46
Hotspot Wizard	Design of small-but-smart libraries and point mutations to improve stability, activity, specificity and selectivity	Webserver https://loschmidt.chemi.muni.cz/hotspotwizard/	87

RosettaRemodel[88] allows users to design novel structures and its functionalities included domain insertions and motif grafting as well as a set of validation tools. Similarly, Rosetta FunFolDes[89] provides a method for grafting functional regions onto proteins. Capabilities of the Rosetta suite were used to support the effort focused on the engineering of subunit-fused nitrile hydratase. There, the utilization of Rosetta resulted in an increase in thermostability and catalytic efficiency[29] by the introduction of single point mutations into a part of the protein that is associated with stability.

Hotspot Wizard[87] is an example of a fully automated web-based multi-purpose tool. It allows users to design small-but-smart libraries solely from the input sequence (**Fig. 7**). The tool can process the sequence to either obtain the structure file from the PDB database or construct a homology model using Modeller[90] and I-Tasser.[91] Furthermore, it contains an internal validation module that serves to alert the user about the low-quality regions of the structure that should be used for modeling only with great caution. Subsequently, the structure is processed by a variety of tools to determine functional and stability (back-to-consensus and structure-based) hot spots as well as correlation-derived hot spots. The resulting hot spots can be then used to generate smart libraries of variants. The software provides information on the putative effects of the mutations in terms of mutations being stabilizing or destabilizing. After the mutations are selected, the user can determine precisely which amino acids should be tested on which position and is presented with the list of necessary degenerated codons together with the report on presence of undesired amino acids in the

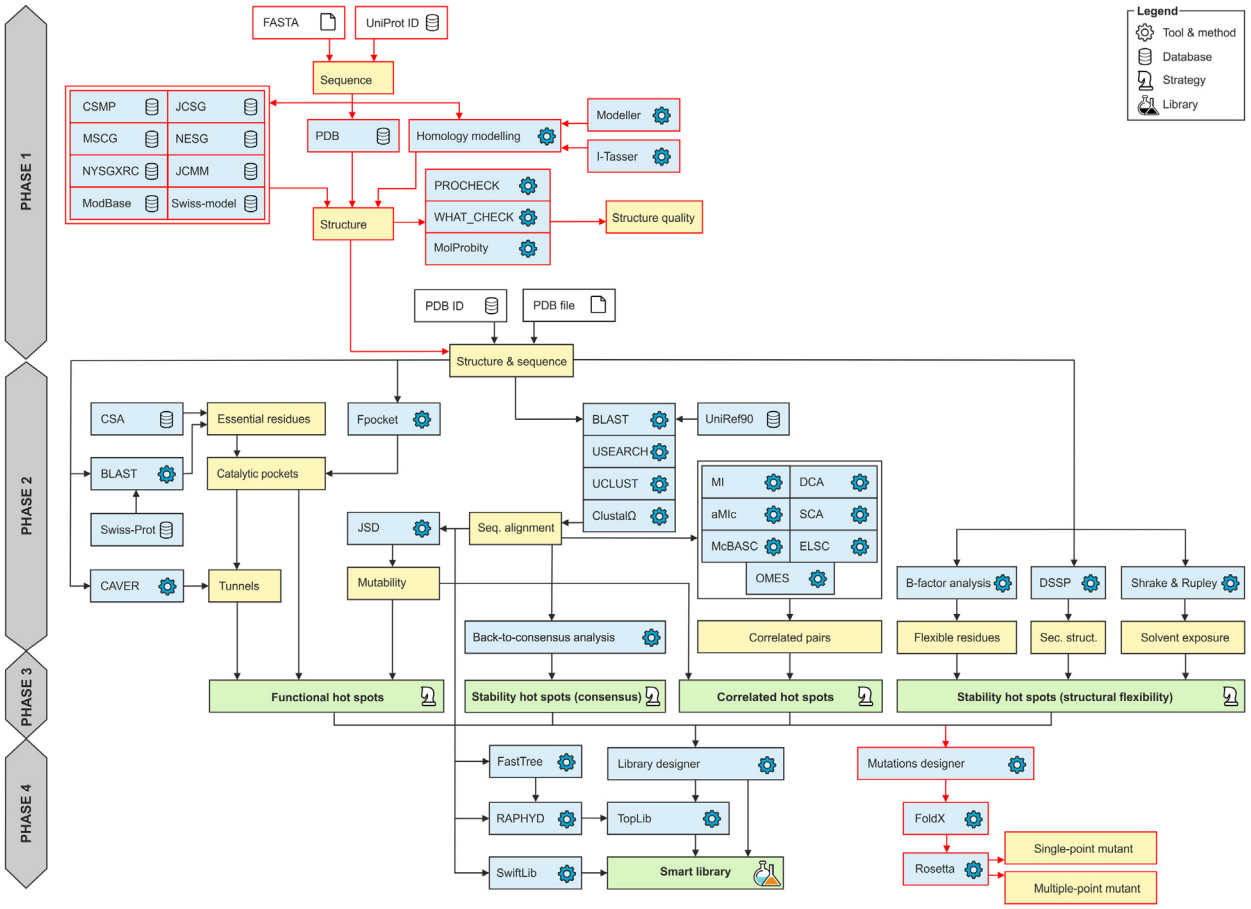


Fig. 7 Workflow diagram of HotSpot Wizard 3.0. The workflow consists of four phases: (1) construction of a model of a structure, (2) annotation of a protein, (3) identification of mutagenesis hot spots and (4) design of mutations and a smart library. Phase 1 is applied only when a sequence is submitted as the input information and this option was added to version 3.0. Reproduced with permission from Ref. 87, Oxford University Press.

library as well as the frequencies stemming from the chosen codon. The web-server contains also molecular docking and functionalities allowing visualization of tunnels and cavities. This vast functionalities are enable by plethora of incorporated tools. For example, aforementioned Caver is used to detect tunnels, and Rosetta is used for design of single and multiple point mutants. The performance of the server was evaluated experimentally by saturation mutagenesis of the residue L177 located at the entrance tunnel of haloalkane dehalogenase LinB.

Over the years, HotSpot Wizard was used in a large number of research projects. Here, we list three biocatalysis-related publications that demonstrate software's versatility. Its ability to calculate cavity sizes was utilized for the development of a hybrid biocatalyst by anchoring a Grubbs-Hoveyda catalyst inside the β-barrel protein scaffold. The resulting catalyst was subsequently employed for ring-opening, ring-closing as well as cross metathesis.[92] Moreover, it was also used to detect the active site in xylanase from *Talaromyces leycettanus* during the study focused on improving its thermostability.[93] Moreover, together with Caver, it was used to engineer ketoreductase from *Chryseobacterium* sp. CA49 (ChKRED20) to enable the reduction of arylketoesters.[94]

Protein visualization is an integral part of protein design. Therefore, it would be a mistake to not mention visualization tools, with the most widely used examples such as **PyMol**,[95] **UCSF Chmiera**[96] and **Yasara**.[97] These tools provide an indispensable ability to look inside the protein structure, annotate and color residues, measure distances and angles, mutate residues, etc. Moreover, they provide an environment to further processing of 3D structures. For example, Chimera and Yasara offer docking functionalities, and Yasara is known for its molecular dynamics functionalities. Chimera is for free, while PyMol and Yasara require a purchase of commercial licenses to unlock the advanced functionalities.

Summary and Outlook

This chapter introduces over 40 software tools that in our opinion constitute a state-of-the-art toolbox for rational protein design. We have focused on the tools that span from the conception of the biocatalytic process to the improvements of the catalysts by optimization of activity, specificity or stability, therefore, we have listed some notable achievements while trying to adhere to biocatalytic applications. However, we have to admit that the utilization of these tools for computational enzyme design is not as widespread as is desirable. Whether it owes to the prejudice among experimentalists or the low success rate of the tools is unclear.

Even though the *in silico* tools have undergone tremendous improvement in the last few decades and they have contributed significantly to the growth of the field, several challenges have to be addressed to catch up with current experimental trends. For example, there is a growing interest in *de novo* protein design,[98] which would provide practically useful biocatalysts embedded into novel folds. It might be caused by a vast amount of protein sequences that are available in genomic databases that have not been studied experimentally, which comprise a limited number of protein folds.

Related to *de novo* protein design is the development of new-to-nature reactions.[99] These reactions are designed by humans but they do not have their nature-born counterparts. The alteration of enzymes to perform new-to-nature reactions was perfectly demonstrated by the Nobel Prize winner Frances Arnold and coworkers in the seminal publications that described the power of directed evolution approaches by modification of heme-containing enzymes for the formation of carbon-boron[100] and carbon-silicon bonds.[101] We envision computational tools that will be able to design active sites based on the spatial configurations of transition states and reaction intermediates, and subsequently graft them onto already known protein folds or come up with entirely novel ones.

Moreover, there has been a large increase in methods that aim to expand the available genetic code.[102] We expect that currently available tools will be forced to adapt to the increased use of non-natural amino acids[103] and new co-factors.[104] This particular technology holds a great promise to expand the protein building blocks from twenty to possibly hundreds. These can be used in several ways, e.g., different amino acids can unlock novel reaction manifolds due to the possibility to precisely tune the properties of the active site and by providing diverse functional groups.

An essential part of this book chapter is focused on the enhancement of protein expression and solubility. Prediction of stabilizing mutations can be considered a solved problem. There is a great number of computational approaches, which provide sufficiently accurate predictions to correctly identify mutations improving protein thermodynamic stability. The most successful are hybrid approaches, combining predictions from fundamentally different approaches, like protein evolution, molecular mechanics, neural networks, etc. Ancestral reconstruction based on phylogenetic analysis and reconstruction of ancient proteins represents another powerful approach for obtaining robust biocatalyst.

Another layer of complexity is represented by the design of enzymes stable in organic solvents. This is because certain chemical reactions are sensitive to the presence of water. On the other hand, enzymes prefer the water environment to that of apolar organic solvents. Indeed, there are experimental workarounds[86] that enable the utilization of enzymes in organic solvents, but the possibility of direct use would open novel biocatalytic avenues. The discussion of whether it would be more beneficial to perform all necessary reaction steps in water or to adapt enzymes to the organic solvents is an interesting one,[105] especially taking into account the various solvents[106] and their possible impact on yield, toxicity, and activity.[86]

Web-based software tools are becoming very popular in the community of protein engineers. These tools do not require any installations and are easy to use. Software developers usually thoroughly validate integrated tools and identify parameters of calculation leading to optimal results. The input data can be inserted via a graphical interface or uploaded, calculations are conducted using the infrastructure of software developers and the outputs are visualized directly in the web browsers. The general trend is to

build sophisticated computational workflows composed of multiple steps and a wide range of algorithms. These features make the use of web tools very efficient and a preferred option over stand-alone software tools.

Machine learning penetrates many scientific fields and computational enzyme design is no exception. Biology became rich in data thanks to technological breakthroughs in DNA sequencing, determination of protein structures by X-ray crystallography, NMR spectrometry and cryo-electron microscopy, and large-scale proteomics by mass spectrometry. Systematically collected and curated data stored in public databases provided valuable sources for supervised and well as unsupervised learning. General and explainable models complement more traditional approaches based on molecular mechanics, molecular dynamics, molecular docking and quantum mechanics. We expect that marriage of modeling approaches with statistical analyses will become a norm in the coming years.

Acknowledgments

The authors would like to express their gratitude to the Czech Ministry of Education (INBIO - CZ.02.1.01/0.0/0.0/16_026/0008451, RECETOX research infrastructure no. LM2018121, ESFRI ELIXIR - LM2018131) for financial support. This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857560. This publication reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains. Jan Vilim is supported by the grant Postdoc2MUNI CZ.02.2.69/0.0/0.0/18_053/0016952.

References

1. Bell, E. L.; Finnigan, W.; France, S. P.; Green, A. P.; Hayes, M. A.; Hepworth, L. J.; Lovelock, S. L.; Niikura, H.; Osuna, S.; Romero, E.; Ryan, K. S.; Turner, N. J.; Flitsch, S. L. Biocatalysis. *Nat. Rev. Methods Prim.* **2021,** *1* (1), 1–21. https://doi.org/10.1038/s43586-021-00044-z.
2. Kazlauskas, R. J.; Bornscheuer, U. T. Finding Better Protein Engineering Strategies. *Nat. Chem. Biol.* **2009,** *5* (8), 526–529. https://doi.org/10.1038/nchembio0809-526.
3. Planas-Iglesias, J.; Marques, S. M.; Pinto, G. P.; Musil, M.; Stourac, J.; Damborsky, J.; Bednar, D. Computational Design of Enzymes for Biotechnological Applications. *Biotechnol. Adv.* **2021,** *47*. https://doi.org/10.1016/j.biotechadv.2021.107696.
4. Marques, S. M.; Planas-Iglesias, J.; Damborsky, J. Web-Based Tools for Computational Enzyme Design. *Curr. Opin. Struct. Biol.* **2021,** *69*, 19–34. https://doi.org/10.1016/j.sbi.2021.01.010.
5. Vasina, M.; Velecký, J.; Planas-Iglesias, J.; Marques, S. M.; Skarupova, J.; Damborsky, J.; Bednar, D.; Mazurenko, S.; Prokop, Z. Tools for Computational Design and High-Throughput Screening of Therapeutic Enzymes. *Adv. Drug Deliv. Rev.* **2022,** *183*, 114143. https://doi.org/10.1016/j.addr.2022.114143.
6. Li, Y.; Wang, S.; Umarov, R.; Xie, B.; Fan, M.; Li, L.; Gao, X. DEEPre: Sequence-Based Enzyme EC Number Prediction by Deep Learning. *Bioinformatics* **2018,** *34* (5), 760–769. https://doi.org/10.1093/bioinformatics/btx680.
7. Dalkiran, A.; Rifaioglu, A. S.; Martin, M. J.; Cetin-Atalay, R.; Atalay, V.; Doğan, T. ECPred: A Tool for the Prediction of the Enzymatic Functions of Protein Sequences Based on the EC Nomenclature. *BMC Bioinform.* **2018,** *19* (1), 1–13. https://doi.org/10.1186/s12859-018-2368-y.
8. Memon, S. A.; Khan, K. A.; Naveed, H. HECNet: A Hierarchical Approach to Enzyme Function Classification Using a Siamese Triplet Network. *Bioinformatics* **2020,** *36* (17), 4583–4589. https://doi.org/10.1093/bioinformatics/btaa536.
9. Khan, K. A.; Memon, S. A.; Naveed, H. A Hierarchical Deep Learning Based Approach for Multi-Functional Enzyme Classification. *Protein Sci.* **2021,** *30* (9), 1935–1945. https://doi.org/10.1002/pro.4146.
10. Zhang, T.; Tian, Y.; Yuan, L.; Chen, F.; Ren, A.; Hu, Q. N. Bio2Rxn: Sequence-Based Enzymatic Reaction Predictions by a Consensus Strategy. *Bioinformatics* **2020,** *36* (11), 3600–3601. https://doi.org/10.1093/bioinformatics/btaa135.
11. Hon, J.; Borko, S.; Stourac, J.; Prokop, Z.; Zendulka, J.; Bednar, D.; Martinek, T.; Damborsky, J. EnzymeMiner: Automated Mining of Soluble Enzymes With Diverse Structures, Catalytic Properties and Stabilities. *Nucleic Acids Res.* **2021,** *48* (W1), W104–W109. https://doi.org/10.1093/NAR/GKAA372.
12. Toti, D.; Viet Hung, L.; Tortosa, V.; Brandi, V.; Polticelli, F. LIBRA-WA: A Web Application for Ligand Binding Site Detection and Protein Function Recognition. *Bioinformatics* **2018,** *34* (5), 878–880. https://doi.org/10.1093/bioinformatics/btx715.
13. Viet Hung, L.; Caprari, S.; Bizai, M.; Toti, D.; Polticelli, F. LIBRA: LIgand Binding Site Recognition Application. *Bioinformatics* **2015,** *31* (24), 4020–4022. https://doi.org/10.1093/bioinformatics/btv489.
14. Angles, R.; Arenas-Salinas, M.; García, R.; Reyes-Suarez, J. A.; Pohl, E. GSP4PDB: A Web Tool to Visualize, Search and Explore Protein-Ligand Structural Patterns. *BMC Bioinform.* **2020,** *21*. https://doi.org/10.1186/s12859-020-3352-x.
15. Hadadi, N.; MohammadiPeyhani, H.; Miskovic, L.; Seijo, M.; Hatzimanikatis, V. Enzyme Annotation for Orphan and Novel Reactions Using Knowledge of Substrate Reactive Sites. *Proc. Natl. Acad. Sci. U. S. A.* **2019,** *116* (15), 7298–7307. https://doi.org/10.1073/pnas.1818877116.
16. Cornish-Bowden, A. Current IUBMB Recommendations on Enzyme Nomenclature and Kinetics. *Perspect. Sci.* **2014,** *1* (1–6), 74–87. https://doi.org/10.1016/j.pisc.2014.02.006.
17. Vanacek, P.; Sebestova, E.; Babkova, P.; Bidmanova, S.; Daniel, L.; Dvorak, P.; Stepankova, V.; Chaloupkova, R.; Brezovsky, J.; Prokop, Z.; Damborsky, J. Exploration of Enzyme Diversity by Integrating Bioinformatics With Expression Analysis and Biochemical Characterization. *ACS Catal.* **2018,** *8* (3), 2402–2412. https://doi.org/10.1021/acscatal.7b03523.
18. Wlodawer, A.; Minor, W.; Dauter, Z.; Jaskolski, M. Protein Crystallography for Aspiring Crystallographers or How to Avoid Pitfalls and Traps in Macromolecular Structure Determination. *FEBS J.* **2013,** *280* (22), 5705–5736. https://doi.org/10.1111/febs.12495.
19. McPherson, A.; Gavira, J. A. Introduction to Protein Crystallization. *Acta Crystallogr. Sect. F Struct. Biol. Commun.* **2014,** *70* (1), 2–20. https://doi.org/10.1107/S2053230X13033141.
20. Hu, Y.; Cheng, K.; He, L.; Zhang, X.; Jiang, B.; Jiang, L.; Li, C.; Wang, G.; Yang, Y.; Liu, M. NMR-Based Methods for Protein Analysis. *Anal. Chem.* **2021,** *93* (4), 1866–1879. https://doi.org/10.1021/acs.analchem.0c03830.
21. Carroni, M.; Saibil, H. R. Cryo Electron Microscopy to Determine the Structure of Macromolecular Complexes. *Methods* **2016,** *95* (2016), 78–85. https://doi.org/10.1016/j.ymeth.2015.11.023.
22. Murata, K.; Wolf, M. Cryo-Electron Microscopy for Structural Analysis of Dynamic Biological Macromolecules. *Biochim. Biophys. Acta* **2018,** *1862* (2), 324–334. https://doi.org/10.1016/j.bbagen.2017.07.020.
23. Kuhlman, B.; Bradley, P. Advances in Protein Structure Prediction and Design. *Nat. Rev. Mol. Cell Biol.* **2019,** *20* (11), 681–697. https://doi.org/10.1038/s41580-019-0163-x.

24. Hameduh, T.; Haddad, Y.; Adam, V.; Heger, Z. Homology Modeling in the Time of Collective and Artificial Intelligence. *Comput. Struct. Biotechnol. J.* **2020,** *18,* 3494–3506. https://doi.org/10.1016/j.csbj.2020.11.007.
25. Callaway, E. "It Will Change Everything": DeepMind's AI Makes Gigantic Leap in Solving Protein Structures. *Nature* **2020,** *588* (7837), 203–204. https://doi.org/10.1038/d41586-020-03348-4.
26. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction With AlphaFold. *Nature* **2021,** *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.
27. Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Dustin Schaeffer, R.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; Van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Christopher Garcia, K.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network. *Science* **2021,** *373* (6557), 871–876. https://doi.org/10.1126/science.abj8754.
28. Moore, P. B.; Hendrickson, W. A.; Henderson, R.; Brunger, A. T. The Protein-Folding Problem: Not Yet Solved. *Science* **2022,** *375* (6580), 507. https://doi.org/10.1126/science.abn9422.
29. Xia, Y.; Cui, W.; Cheng, Z.; Peplowski, L.; Liu, Z.; Kobayashi, M.; Zhou, Z. Improving the Thermostability and Catalytic Efficiency of the Subunit-Fused Nitrile Hydratase by Semi-rational Engineering. *ChemCatChem* **2018,** *10* (6), 1370–1375. https://doi.org/10.1002/cctc.201701374.
30. Pagadala, N. S.; Syed, K.; Tuszynski, J. Software for Molecular Docking: A Review. *Biophys. Rev.* **2017,** *9* (2), 91–102. https://doi.org/10.1007/s12551-016-0247-1.
31. Serrano Cardona, L.; Muñoz Mata, E. Paraninfo Digital. *Early Hum. Dev.* **2013,** *83* (1), 1–11. https://doi.org/10.1016/j.earlhumdev.2006.05.022.
32. Goodsell, D. S.; Sanner, M. F.; Olson, A. J.; Forli, S. The AutoDock Suite at 30. *Protein Sci.* **2021,** *30* (1), 31–43. https://doi.org/10.1002/pro.3934.
33. Eberhardt, J.; Santos-Martins, D.; Tillack, A. F.; Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J. Chem. Inf. Model.* **2021,** *61* (8), 3891–3898. https://doi.org/10.1021/acs.jcim.1c00203.
34. Lyskov, S.; Chou, F. C.; Conchúir, S.Ó.; Der, B. S.; Drew, K.; Kuroda, D.; Xu, J.; Weitzner, B. D.; Renfrew, P. D.; Sripakdeevong, P.; Borgo, B.; Havranek, J. J.; Kuhlman, B.; Kortemme, T.; Bonneau, R.; Gray, J. J.; Das, R. Serverification of Molecular Modeling Applications: The Rosetta Online Server That Includes Everyone (ROSIE). *PLoS One* **2013,** *8* (5), 5–7. https://doi.org/10.1371/journal.pone.0063906.
35. Khersonsky, O.; Lipsh, R.; Avizemer, Z.; Ashani, Y.; Goldsmith, M.; Leader, H.; Dym, O.; Rogotner, S.; Trudeau, D. L.; Prilusky, J.; Amengual-Rigo, P.; Guallar, V.; Tawfik, D. S.; Fleishman, S. J. Automated Design of Efficient and Functionally Diverse Enzyme Repertoires. *Mol. Cell* **2018,** *72* (1), 178–186.e5. https://doi.org/10.1016/j.molcel.2018.08.033.
36. Fassio, A. V.; Santos, L. H.; Silveira, S. A.; Ferreira, R. S.; De Melo-Minardi, R. C. NAPOLI: A Graph-Based Strategy to Detect and Visualize Conserved Protein-Ligand Interactions in Large-Scale. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2020,** *17* (4), 1317–1328. https://doi.org/10.1109/TCBB.2019.2892099.
37. Lee, P. H.; Kuo, K. L.; Chu, P. Y.; Liu, E. M.; Lin, J. H. SLITHER: A Web Server for Generating Contiguous Conformations of Substrate Molecules Entering Into Deep Active Sites of Proteins or Migrating Through Channels in Membrane Transporters. *Nucleic Acids Res.* **2009,** *37* (Suppl. 2), 559–564. https://doi.org/10.1093/nar/gkp359.
38. Devaurs, D.; Bouard, L.; Vaisset, M.; Zanon, C.; Al-Bluwi, I.; Iehl, R.; Siméon, T.; Cortés, J. MoMA-LigPath: A Web Server to Simulate Protein-Ligand Unbinding. *Nucleic Acids Res.* **2013,** *41* (Web Server issue), 297–302. https://doi.org/10.1093/nar/gkt380.
39. Sánchez-Aparicio, J. E.; Sciortino, G.; Herrmannsdoerfer, D. V.; Chueca, P. O.; Pedregal, J. R. G.; Maréchal, J. D. GPathFinder: Identification of Ligand-Binding Pathways by a Multi-objective Genetic Algorithm. *Int. J. Mol. Sci.* **2019,** *20* (13). https://doi.org/10.3390/ijms20133155.
40. Stourac, J.; Vavra, O.; Kokkonen, P.; Filipovic, J.; Pinto, G.; Brezovsky, J.; Damborsky, J.; Bednar, D. Caver Web 1.0: Identification of Tunnels and Channels in Proteins and Analysis of Ligand Transport. *Nucleic Acids Res.* **2019,** *47* (W1), W414–W422. https://doi.org/10.1093/nar/gkz378.
41. Aggarwal, R.; Gupta, A.; Chelur, V.; Jawahar, C. V.; Priyakumar, U. D. DeepPocket: Ligand Binding Site Detection and Segmentation Using 3D Convolutional Neural Networks. *J. Chem. Inf. Model.* **2021.** https://doi.org/10.1021/acs.jcim.1c00799.
42. Marchand, J. R.; Pirard, B.; Ertl, P.; Sirockin, F. CAVIAR: A Method for Automatic Cavity Detection, Description and Decomposition Into Subcavities. *J. Comput. Aided Mol. Des.* **2021,** *35* (6), 737–750. https://doi.org/10.1007/s10822-021-00390-w.
43. Karami, Y.; Rey, J.; Postic, G.; Murail, S.; Tufféry, P.; De Vries, S. J. DaReUS-Loop: A Web Server to Model Multiple Loops in Homology Models. *Nucleic Acids Res.* **2019,** *47* (W1), W423–W428. https://doi.org/10.1093/nar/gkz403.
44. Planas-Iglesias, J.; Opaleny, F.; Ulbrich, P.; Stourac, J.; Sanusi, Z.; Pinto, G. P.; Schenkmayerova, A.; Byska, J.; Damborsky, J.; Kozlikova, B.; Bednar, D. LoopGrafter: A Web Tool for Transplanting Dynamical Loops for Protein Engineering. *Nucleic Acids Res.* **2022.** https://doi.org/10.1093/nar/gkac249.
45. Wong, K. M.; Tai, H. K.; Siu, S. W. I. GWOVina: A Grey Wolf Optimization Approach to Rigid and Flexible Receptor Docking. *Chem. Biol. Drug Des.* **2021,** *97* (1), 97–110. https://doi.org/10.1111/cbdd.13764.
46. Leman, J. K.; Weitzner, B. D.; Lewis, S. M.; Adolf-Bryfogle, J.; Alam, N.; Alford, R. F.; Aprahamian, M.; Baker, D.; Barlow, K. A.; Barth, P.; Basanta, B.; Bender, B. J.; Blacklock, K.; Bonet, J.; Boyken, S. E.; Bradley, P.; Bystroff, C.; Conway, P.; Cooper, S.; Correia, B. E.; Coventry, B.; Das, R.; De Jong, R. M.; DiMaio, F.; Dsilva, L.; Dunbrack, R.; Ford, A. S.; Frenz, B.; Fu, D. Y.; Geniesse, C.; Goldschmidt, L.; Gowthaman, R.; Gray, J. J.; Gront, D.; Guffy, S.; Horowitz, S.; Huang, P. S.; Huber, T.; Jacobs, T. M.; Jeliazkov, J. R.; Johnson, D. K.; Kappel, K.; Karanicolas, J.; Khakzad, H.; Khar, K. R.; Khare, S. D.; Khatib, F.; Khramushin, A.; King, I. C.; Kleffner, R.; Koepnick, B.; Kortemme, T.; Kuenze, G.; Kuhlman, B.; Kuroda, D.; Labonte, J. W.; Lai, J. K.; Lapidoth, G.; Leaver-Fay, A.; Lindert, S.; Linsky, T.; London, N.; Lubin, J. H.; Lyskov, S.; Maguire, J.; Malmström, L.; Marcos, E.; Marcu, O.; Marze, N. A.; Meiler, J.; Moretti, R.; Mulligan, V. K.; Nerli, S.; Norn, C.; Ó'Conchúir, S.; Ollikainen, N.; Ovchinnikov, S.; Pacella, M. S.; Pan, X.; Park, H.; Pavlovicz, R. E.; Pethe, M.; Pierce, B. G.; Pilla, K. B.; Raveh, B.; Renfrew, P. D.; Burman, S. S. R.; Rubenstein, A.; Sauer, M. F.; Scheck, A.; Schief, W.; Schueler-Furman, O.; Sedan, Y.; Sevy, A. M.; Sgourakis, N. G.; Shi, L.; Siegel, J. B.; Silva, D. A.; Smith, S.; Song, Y.; Stein, A.; Szegedy, M.; Teets, F. D.; Thyme, S. B.; Wang, R. Y. R.; Watkins, A.; Zimmerman, L.; Bonneau, R. Macromolecular Modeling and Design in Rosetta: Recent Methods and Frameworks. *Nat. Methods* **2020,** *17* (7), 665–680. https://doi.org/10.1038/s41592-020-0848-2.
47. Wang, J.; Liang, J.; Chen, L.; Zhang, W.; Kong, L.; Peng, C.; Su, C.; Tang, Y.; Deng, Z.; Wang, Z. Structural Basis for the Biosynthesis of Lovastatin. *Nat. Commun.* **2021,** *12* (1), 1–10. https://doi.org/10.1038/s41467-021-21174-8.
48. Nam, D.; Tinoco, A.; Shen, Z.; Adukure, R. D.; Sreenilayam, G.; Khare, S. D.; Fasan, R. Enantioselective Synthesis of α-Trifluoromethyl Amines Via Biocatalytic N-H Bond Insertion With Acceptor-Acceptor Carbene Donors. *J. Am. Chem. Soc.* **2022,** *144* (6), 2590–2602. https://doi.org/10.1021/jacs.1c10750.
49. Bengel, L. L.; Aberle, B.; Egler-Kemmerer, A. N.; Kienzle, S.; Hauer, B.; Hammer, S. C. Engineered Enzymes Enable Selective N-Alkylation of Pyrazoles With Simple Haloalkanes. *Angew. Chem. Int. Ed.* **2021,** *60* (10), 5554–5560. https://doi.org/10.1002/anie.202014239.
50. Chang, D. T. H.; Oyang, Y. J.; Lin, J. H. MEDock: A Web Server for Efficient Prediction of Ligand Binding Sites Based on a Novel Optimization Algorithm. *Nucleic Acids Res.* **2005,** *33* (Suppl. 2). https://doi.org/10.1093/nar/gki586.
51. Streltsov, V. A.; Luang, S.; Peisley, A.; Varghese, J. N.; Ketudat Cairns, J. R.; Fort, S.; Hijnen, M.; Tvaroška, I.; Ardá, A.; Jiménez-Barbero, J.; Alfonso-Prieto, M.; Rovira, C.; Mendoza, F.; Tiessler-Sala, L.; Sánchez-Aparicio, J. E.; Rodríguez-Guerra, J.; Lluch, J. M.; Maréchal, J. D.; Masgrau, L.; Hrmova, M. Discovery of Processive Catalysis by an Exo-hydrolase With a Pocket-Shaped Active Site. *Nat. Commun.* **2019,** *10* (1), 1–17. https://doi.org/10.1038/s41467-019-09691-z.
52. Chovancova, E.; Pavelka, A.; Benes, P.; Strnad, O.; Brezovsky, J.; Kozlikova, B.; Gora, A.; Sustr, V.; Klvana, M.; Medek, P.; Biedermannova, L.; Sochor, J.; Damborsky, J. CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures. *PLoS Comput. Biol.* **2012,** *8* (10). https://doi.org/10.1371/journal.pcbi.1002708.
53. Vavra, O.; Filipovic, J.; Plhak, J.; Bednar, D.; Marques, S. M.; Brezovsky, J.; Stourac, J.; Matyska, L.; Damborsky, J. CaverDock: A Molecular Docking-Based Tool to Analyse Ligand Transport Through Protein Tunnels and Channels. *Bioinformatics* **2019,** *35* (23), 4986–4993. https://doi.org/10.1093/bioinformatics/btz386.

54. Meng, S.; Ji, Y.; Liu, L.; Davari, M. D.; Schwaneberg, U. Modulating the Coupling Efficiency of P450 BM3 by Controlling Water Diffusion Through Access Tunnel Engineering. *ChemSusChem* **2022,** *15* (9). https://doi.org/10.1002/cssc.202102434.
55. Godehard, S. P.; Badenhorst, C. P. S.; Müller, H.; Bornscheuer, U. T. Protein Engineering for Enhanced Acyltransferase Activity, Substrate Scope, and Selectivity of the Mycobacterium Smegmatis Acyltransferase MsAcT. *ACS Catal.* **2020,** *10* (14), 7552–7562. https://doi.org/10.1021/acscatal.0c01767.
56. Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinform.* **2009,** *10,* 1–11. https://doi.org/10.1186/1471-2105-10-168.
57. Tawfik, D. S. Loop Grafting and the Origins of Enzyme Species. *Science* **2006,** *311* (5760), 475–476. https://doi.org/10.1126/science.1123883.
58. Park, H. S.; Nam, S. H.; Lee, J. K.; Yoon, C. N.; Mannervik, B.; Benkovic, S. J.; Kim, H. S. Design and Evolution of New Catalytic Activity With an Existing Protein Scaffold. *Science* **2006,** *311* (5760), 535–538. https://doi.org/10.1126/science.1118953.
59. Dodani, S. C.; Kiss, G.; Cahn, J. K. B.; Su, Y.; Pande, V. S.; Arnold, F. H. Discovery of a Regioselectivity Switch in Nitrating P450s Guided by Molecular Dynamics Simulations and Markov Models. *Nat. Chem.* **2016,** *8* (5), 419–425. https://doi.org/10.1038/nchem.2474.
60. Karami, Y.; Guyon, F.; De Vries, S.; Tufféry, P. DaReUS-Loop: Accurate Loop Modeling Using Fragments from Remote or Unrelated Proteins. *Sci. Rep.* **2018,** *8* (1), 1–12. https://doi.org/10.1038/s41598-018-32079-w.
61. Schymkowitz, J.; Borg, J.; Stricher, F.; Nys, R.; Rousseau, F.; Serrano, L. The FoldX Web Server: An Online Force Field. *Nucleic Acids Res.* **2005,** *33* (Suppl. 2), 382–388. https://doi.org/10.1093/nar/gki387.
62. Wijma, H. J.; Fürst, M. J. L. J.; Janssen, D. B. A Computational Library Design Protocol for Rapid Improvement of Protein Stability: FRESCO. *Methods Mol. Biol.* **2018,** *1685* (6), 69–85. https://doi.org/10.1007/978-1-4939-7366-8_5.
63. Goldenzweig, A.; Goldsmith, M.; Hill, S. E.; Gertman, O.; Laurino, P.; Ashani, Y.; Dym, O.; Unger, T.; Albeck, S.; Prilusky, J.; Lieberman, R. L.; Aharoni, A.; Silman, I.; Sussman, J. L.; Tawfik, D. S.; Fleishman, S. J. Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Mol. Cell* **2016,** *63* (2), 337–346. https://doi.org/10.1016/j.molcel.2016.06.012.
64. Banerjee, A.; Mitra, P. Estimating the Effect of Single-Point Mutations on Protein Thermodynamic Stability and Analyzing the Mutation Landscape of the P53 Protein. *J. Chem. Inf. Model.* **2020,** *60* (6), 3315–3323. https://doi.org/10.1021/acs.jcim.0c00256.
65. Musil, M.; Stourac, J.; Bendl, J.; Brezovsky, J.; Prokop, Z.; Zendulka, J.; Martinek, T.; Bednar, D.; Damborsky, J. FireProt: Web Server for Automated Design of Thermostable Proteins. *Nucleic Acids Res.* **2017,** *45* (W1), W393–W399. https://doi.org/10.1093/nar/gkx285.
66. Rodrigues, C. H. M.; Pires, D. E. V.; Ascher, D. B. DynaMut: Predicting the Impact of Mutations on Protein Conformation, Flexibility and Stability. *Nucleic Acids Res.* **2018,** *46* (W1), W350–W355. https://doi.org/10.1093/nar/gky300.
67. Contessoto, V. G.; de Oliveira, V. M.; Fernandes, B. R.; Slade, G. G.; Leite, V. B. P. TKSA-MC: A Web Server for Rational Mutation Through the Optimization of Protein Charge Interactions. *Proteins Struct. Funct. Bioinform.* **2018,** *86* (11), 1184–1188. https://doi.org/10.1002/prot.25599.
68. Gopi, S.; Devanshu, D.; Krishna, P.; Naganathan, A. N. PStab: Prediction of Stable Mutants, Unfolding Curves, Stability Maps and Protein Electrostatic Frustration. *Bioinformatics* **2018,** *34* (5), 875–877. https://doi.org/10.1093/bioinformatics/btx697.
69. Li, B.; Yang, Y. T.; Capra, J. A.; Gerstein, M. B. Predicting Changes in Protein Thermodynamic Stability Upon Point Mutation With Deep 3D Convolutional Neural Networks. *PLoS Comput. Biol.* **2020,** *16* (11), 1–24. https://doi.org/10.1371/journal.pcbi.1008291.
70. Suplatov, D.; Timonina, D.; Sharapova, Y.; Švedas, V. Yosshi: A Web-Server for Disulfide Engineering by Bioinformatic Analysis of Diverse Protein Families. *Nucleic Acids Res.* **2019,** *47* (W1), W308–W314. https://doi.org/10.1093/nar/gkz385.
71. Musil, M.; Khan, R. T.; Beier, A.; Stourac, J.; Konegger, H.; Damborsky, J.; Bednar, D. FireProtASR: A Web Server for Fully Automated Ancestral Sequence Reconstruction. *Brief. Bioinform.* **2021,** *22* (4). https://doi.org/10.1093/bib/bbaa337.
72. Fürst, M. J. L. J.; Boonstra, M.; Bandstra, S.; Fraaije, M. W. Stabilization of Cyclohexanone Monooxygenase by Computational and Experimental Library Design. *Biotechnol. Bioeng.* **2019,** *116* (9), 2167–2177. https://doi.org/10.1002/bit.27022.
73. Meng, Q.; Capra, N.; Palacio, C. M.; Lanfranchi, E.; Otzen, M.; Van Schie, L. Z.; Rozeboom, H. J.; Thunnissen, A. M. W. H.; Wijma, H. J.; Janssen, D. B. Robust ω-Transaminases by Computational Stabilization of the Subunit Interface. *ACS Catal.* **2020,** *10* (5), 2915–2928. https://doi.org/10.1021/acscatal.9b05223.
74. Barber-Zucker, S.; Mindel, V.; Garcia-Ruiz, E.; Weinstein, J. J.; Alcalde, M.; Fleishman, S. J. Stable and Functionally Diverse Versatile Peroxidases Designed Directly From Sequences. *J. Am. Chem. Soc.* **2022,** *144* (8), 3564–3571. https://doi.org/10.1021/jacs.1c12433.
75. Peleg, Y.; Vincentelli, R.; Collins, B. M.; Chen, K. E.; Livingstone, E. K.; Weeratunga, S.; Leneva, N.; Guo, Q.; Remans, K.; Perez, K.; Bjerga, G. E. K.; Larsen, Ø.; Vaněk, O.; Skořepa, O.; Jacquemin, S.; Poterszman, A.; Kjær, S.; Christodoulou, E.; Albeck, S.; Dym, O.; Ainbinder, E.; Unger, T.; Schuetz, A.; Matthes, S.; Bader, M.; de Marco, A.; Storici, P.; Semrau, M. S.; Stolt-Bergner, P.; Aigner, C.; Suppmann, S.; Goldenzweig, A.; Fleishman, S. J. Community-Wide Experimental Evaluation of the PROSS Stability-Design Method. *J. Mol. Biol.* **2021,** *433* (13). https://doi.org/10.1016/j.jmb.2021.166964.
76. Liu, Y.; Li, Z. Y.; Guo, C.; Cui, C.; Lin, H.; Wu, Z. L. Enhancing the Thermal Stability of Ketoreductase ChKRED12 Using the FireProt Web Server. *Process Biochem.* **2021,** *101,* 207–212. https://doi.org/10.1016/j.procbio.2020.11.018. July 2020.
77. Goldenzweig, A.; Fleishman, S. J. Principles of Protein Stability and Their Application in Computational Design. *Annu. Rev. Biochem.* **2018,** 105–129. https://doi.org/10.1146/annurev-biochem-062917-012102. Department of Biomolecular Sciences, Weizmann Institute of Science, Rehovot, 76100, Israel.
78. Hirose, S.; Noguchi, T. Espresso: A System for Estimating Protein Expression and Solubility in Protein Expression Systems. *Proteomics* **2013,** *13* (9), 1444–1456. https://doi.org/10.1002/pmic.201200175.
79. Madani, M.; Lin, K.; Tarakanova, A. DSResSol: A Sequence-Based Solubility Predictor Created With Dilated Squeeze Excitation Residual Networks. *Int. J. Mol. Sci.* **2021,** *22* (24). https://doi.org/10.3390/ijms222413555.
80. Khurana, S.; Rawi, R.; Kunji, K.; Chuang, G. Y.; Bensmail, H.; Mall, R. DeepSol: A Deep Learning Framework for Sequence-Based Protein Solubility Prediction. *Bioinformatics* **2018,** *34* (15), 2605–2613. https://doi.org/10.1093/bioinformatics/bty166.
81. Hon, J.; Marusiak, M.; Martinek, T.; Kunka, A.; Zendulka, J.; Bednar, D.; Damborsky, J. SoluProt: Prediction of Soluble Protein Expression in *Escherichia Coli*. *Bioinformatics* **2021,** *37* (1), 23–28. https://doi.org/10.1093/bioinformatics/btaa1102.
82. Hou, Q.; Kwasigroch, J. M.; Rooman, M.; Pucci, F. SOLart: A Structure-Based Method to Predict Protein Solubility and Aggregation. *Bioinformatics* **2020,** *36* (5), 1445–1452. https://doi.org/10.1093/bioinformatics/btz773.
83. Bhandari, B. K.; Gardner, P. P.; Lim, C. S. Solubility-Weighted Index: Fast and Accurate Prediction of Protein Solubility. *Bioinformatics* **2020,** *36* (18), 4691–4698. https://doi.org/10.1093/bioinformatics/btaa578.
84. Kuriata, A.; Iglesias, V.; Pujols, J.; Kurcinski, M.; Kmiecik, S.; Ventura, S. Aggrescan3D (A3D) 2.0: Prediction and Engineering of Protein Solubility. *Nucleic Acids Res.* **2019,** *47* (W1), W300–W307. https://doi.org/10.1093/nar/gkz321.
85. Tsumoto, K.; Ejima, D.; Kumagai, I.; Arakawa, T. Practical Considerations in Refolding Proteins From Inclusion Bodies. *Protein Expr. Purif.* **2003,** *28* (1), 1–8. https://doi.org/10.1016/S1046-5928(02)00641-1.
86. Stepankova, V.; Bidmanova, S.; Koudelakova, T.; Prokop, Z.; Chaloupkova, R.; Damborsky, J. Strategies for Stabilization of Enzymes in Organic Solvents. *ACS Catal.* **2013,** *3* (12), 2823–2836. https://doi.org/10.1021/cs400684x.
87. Sumbalova, L.; Stourac, J.; Martinek, T.; Bednar, D.; Damborsky, J. HotSpot Wizard 3.0: Web Server for Automated Design of Mutations and Smart Libraries Based on Sequence Input Information. *Nucleic Acids Res.* **2018,** *46* (W1), W356–W362. https://doi.org/10.1093/nar/gky417.
88. Huang, P. S.; Ban, Y. E. A.; Richter, F.; Andre, I.; Vernon, R.; Schief, W. R.; Baker, D. Rosettaremodel: A Generalized Framework for Flexible Backbone Protein Design. *PLoS One* **2011,** *6* (8). https://doi.org/10.1371/journal.pone.0024109.

89. Bonet, J.; Wehrle, S.; Schriever, K.; Yang, C.; Billet, A.; Sesterhenn, F.; Scheck, A.; Sverrisson, F.; Veselkova, B.; Vollers, S.; Lourman, R.; Villard, M.; Rosset, S.; Krey, T.; Correia, B. E. Rosetta FunFolDes—A General Framework for the Computational Design of Functional Proteins. *PLoS Comput. Biol.* **2018,** *14* (11), 1–30. https://doi.org/10.1371/journal.pcbi.1006623.
90. Webb, B.; Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinform.* **2016,** *54,* 5.6.1–5.6.37. https://doi.org/10.1002/cpbi.3.
91. Yang, J.; Yan, R.; Roy, A.; Xu, D.; Poisson, J.; Zhang, Y. The I-TASSER Suite: Protein Structure and Function Prediction. *Nat. Methods* **2014,** *12* (1), 7–8. https://doi.org/10.1038/nmeth.3213.
92. Grimm, A. R.; Sauer, D. F.; Davari, M. D.; Zhu, L.; Bocola, M.; Kato, S.; Onoda, A.; Hayashi, T.; Okuda, J.; Schwaneberg, U. Cavity Size Engineering of a β-Barrel Protein Generates Efficient Biohybrid Catalysts for Olefin Metathesis. *ACS Catal.* **2018,** *8* (4), 3358–3364. https://doi.org/10.1021/acscatal.7b03652.
93. Wang, X.; Ma, R.; Xie, X.; Liu, W.; Tu, T.; Zheng, F.; You, S.; Ge, J.; Xie, H.; Yao, B.; Luo, H. Thermostability Improvement of a Talaromyces Leycettanus Xylanase by Rational Protein Engineering. *Sci. Rep.* **2017,** *7* (1), 1–9. https://doi.org/10.1038/s41598-017-12659-y.
94. Li, T. B.; Zhao, F. J.; Liu, Z.; Jin, Y.; Liu, Y.; Pei, X. Q.; Zhang, Z. G.; Wang, G.; Wu, Z. L. Structure-Guided Engineering of ChKRED20 From *Chryseobacterium* sp. CA49 for Asymmetric Reduction of Aryl Ketoesters. *Enzyme Microb. Technol.* **2019,** *125,* 29–36. https://doi.org/10.1016/j.enzmictec.2019.03.001.
95. Delano, W. L. *The PyMOL Molecular Graphics System*, CCP4 Newsletter On Protein Crystallography, 2002; pp 82–92.
96. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera—A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004,** *25* (13), 1605–1612. https://doi.org/10.1002/jcc.20084.
97. Krieger, E.; Vriend, G. New Ways to Boost Molecular Dynamics Simulations. *J. Comput. Chem.* **2015,** *36* (13), 996–1007. https://doi.org/10.1002/jcc.23899.
98. Reetz, M. Making Enzymes Suitable for Organic Chemistry by Rational Protein Design. *ChemBioChem* **2022,** 202200049. https://doi.org/10.1002/cbic.202200049.
99. Arnold, F. H. Innovation by Evolution: Bringing New Chemistry to Life (Nobel Lecture). *Angew. Chem. Int. Ed.* **2019,** *58* (41), 14420–14426. https://doi.org/10.1002/anie.201907729.
100. Jennifer Kan, S. B.; Huang, X.; Gumulya, Y.; Chen, K.; Arnold, F. H. Genetically Programmed Chiral Organoborane Synthesis. *Nature* **2017,** *552* (7683), 132–136. https://doi.org/10.1038/nature24996.
101. Kan, S. B. J.; Lewis, R. D.; Chen, K.; Arnold, F. H. Directed Evolution of Cytochrome c for Carbon-Silicon Bond Formation: Bringing Silicon to Life. *Science* **2016,** *354* (6315), 1048–1051. https://doi.org/10.1126/science.aah6219.
102. Shandell, M. A.; Tan, Z.; Cornish, V. W. Genetic Code Expansion: A Brief History and Perspective. *Biochemistry* **2021,** *60* (46), 3455–3469. https://doi.org/10.1021/acs.biochem.1c00286.
103. Drienovská, I.; Roelfes, G. Expanding the Enzyme Universe With Genetically Encoded Unnatural Amino Acids. *Nat. Catal.* **2020,** *3* (3), 193–202. https://doi.org/10.1038/s41929-019-0410-8.
104. Guarneri, A.; van Berkel, W. J.; Paul, C. E. Alternative Coenzymes for Biocatalysis. *Curr. Opin. Biotechnol.* **2019,** *60,* 63–71. https://doi.org/10.1016/j.copbio.2019.01.001.
105. Holtmann, D.; Hollmann, F. Is Water the Best Solvent for Biocatalysis? *Mol. Catal.* **2022,** *517* (December 2021), 112035 https://doi.org/10.1016/j.mcat.2021.112035.
106. Alder, C. M.; Hayler, J. D.; Henderson, R. K.; Redman, A. M.; Shukla, L.; Shuster, L. E.; Sneddon, H. F. Updating and Further Expanding GSK's Solvent Sustainability Guide. *Green Chem.* **2016,** *18* (13), 3879–3890. https://doi.org/10.1039/c6gc00611f.