Published in final edited form as:
Nucleic Acids Res. 2022 July 05; 50(W1): W13–W20. doi:10.1093/nar/gkac250.

3DLigandSite: Structure-based prediction of protein-ligand binding sites

Jake E McGreig[1], **Hannah Uri**[1], **Magdalena Antczak**[1], **Michael JE Sternberg**[2], **Martin Michaelis**[1], **Mark N Wass**[1,*]

[1]School of Biosciences, Division of Natural Sciences, University of Kent, Canterbury, Kent, CT2 7NJ, UK

[2]Centre for Integrative Systems Biology and Bioinformatics, Department of Life Sciences, Imperial College London, London, SW7 2AZ, UK

Abstract

3DLigandSite is a web tool for the prediction of ligand-binding sites in proteins. Here, we report a significant update since the first release of 3DLigandSite in 2010. The overall methodology remains the same, with candidate binding sites in proteins inferred using known binding sites in related protein structures as templates. However, the initial structural modelling step now uses the newly available structures from the AlphaFold database or alternatively Phyre2 when AlphaFold structures are not available. Further, a sequence-based search using HHSearch has been introduced to identify template structures with bound ligands that are used to infer the ligand-binding residues in the query protein. Finally, we introduced a machine learning element as the final prediction step, which improves the accuracy of predictions and provides a confidence score for each residue predicted to be part of a binding site. Validation of 3DLigandSite on a set of 6416 binding sites obtained 92% recall at 75% precision for non-metal binding sites and 52% recall at 75% precision for metal binding sites. 3DLigandSite is available at https://www.wass-michaelislab.org/3dligandsite. Users submit either a protein sequence or structure. Results are displayed in multiple formats including an interactive Mol* molecular visualisation of the protein and the predicted binding sites.

Introduction

Elucidation of protein function remains a difficult and important task, with many millions of proteins present in UniProt [1] and only a small fraction of them functionally annotated [2,3], making automated sequence annotation tools essential. Small molecules that bind to proteins are intimately related to protein function; they can be substrates or products of

*To whom correspondence should be addressed. Tel: +44 (0)1227 827626; m.n.wass@kent.ac.uk.

Conflict of Interest
The authors have no conflicts of interest to declare.

Author Contributions
MNW, MM and MJES devised the research. JEM, HU, MA performed the research and analysed data. MNW and JEM wrote manuscript with contributions from all other authors.

Europe PMC Funders Author Manuscripts

an enzyme reaction, cofactors [4] that play an essential role in catalysis or have important structural or regulatory roles [5].

Methods for predicting ligand-binding sites (reviewed in [6]) use a range of different approaches, including sequence conservation [7], structural approaches such as identifying pockets on the protein surface, the combined analysis of sequence and structural information [8], and machine and deep learning [9–16]. 3DLigandSite and methods such as firestar [17], FINDSITE [18,19], COACH-D [20] and FunFOLD2 [21] utilise knowledge of existing binding sites in solved protein structures present in the Protein Data Bank (PDB) [22]. 3DLigandSite, FINDSITE, FunFOLD2 and COACH-D combine the modelling of protein structure with the identification of homologous proteins in the PDB that have ligands bound to them. These binding sites are then used to infer binding sites in the query protein. By contrast, firestar uses FireDB [23], a database of ligand-binding residues extracted from protein structures in the PDB and also catalytic residues extracted from the catalytic site atlas [24].

Here, we present the first major update to the 3DLigandSite web server. 3DLigandSite was first developed in 2010 [25] to automate an approach that was successfully used in the ligand-binding site experiment in the 8th round of the Critical Assessment of protein Structure Prediction (CASP) community experiment [26,27]. Over the last twelve years, 3DLigandSite has become widely used, attracting an average of 125,000 submissions per year, for a diverse range of purposes, including genome annotation [28,29], antiviral screening [30], the analysis of single nucleotide variants associated with disease [31–35], the development of fluorescent sensors [36] and most recently for analysis of SARS coronavirus-2 proteins [37–39]. Over the last three years, 3DLigandSite binding site predictions have been incorporated into the Protein Data Bank in Europe (PDBe; [22]) Knowledgebase [40,41], making binding site predictions for protein structures in the PDBe widely available.

The basic 3DLigandSite algorithm remains the same in the new version, but it now makes use of the latest sequence searching methods and the highly accurate protein 3D structural models available from the AlphaFold Protein Structure Database (AlphaFold DB [42,43]). 3DLigandSite now also incorporates machine learning as the final step in the prediction process, which improves prediction accuracy and associates a confidence score with each individual residue predicted to be part of a binding site. This is combined with a new web server that offers improved functionality for users to investigate the predicted binding sites.

The 3DLigandSite Method

A summary of the 3DLigandSite methodology is outlined in Figure 1. Users submit either a protein sequence in FASTA format or a protein structure in PDB format. Where a sequence is submitted, the PDB is first searched for an existing structure with identical sequence that can be used. Where a match is not found, AlphaFold DB [42] is searched for an existing structural model. Finally, where a model is not available, Phyre2 [44] is used to perform template-based modelling.

The next step focuses on the identification of ligand-bound structures that are homologous to the query protein. Originally, 3DLigandSite used MAMMOTH [45] to perform a structural search of the query structure against a structural library of proteins from the PDB, which was a time-consuming step, typically taking between 40-80 minutes. This has been replaced by a sequence-based search using HHSearch [46] to screen a sequence library of ligand-bound proteins from the PDB (detailed below), which only takes a few minutes to run. All sequence matches with an HHSearch probability score greater than 75 % are retained, and their protein structures are aligned to the query structure using TM-Align [47]. The user can reduce the HHSearch probability cut off if they would like to use less confident matches to the query sequence.

Where matches to the library of ligand-bound proteins are not identified by the sequence-based search, a structural search is performed using TM-Align [47] that retains alignments with a TM-Score of 0.6 or greater. The structural search is also available as an advanced option that users can choose to perform at the time of submission.

The ligands present in the library structures are superimposed onto the query structure by aligning the library structures with the query structure. These ligands are then clustered. Originally, 3DLigandSite used single linkage clustering to cluster ligands, which could result in very large clusters. To avoid this, 3DLigandSite now generates clusters such that 50 % of each ligand must overlap with at least one of the other ligands in the same cluster. This change also required that metal and non-metal ligands are separately clustered given that metal ligands are single ions, while non-metal ligands are larger molecules (e.g. ATP, NAD). Individual predictions of metal and non-metal binding sites are also made for these separate clusters.

The final step of the prediction process is to determine the residues in the protein that are predicted to form the binding site associated with each cluster of ligands. Each cluster may contain multiple different ligands or many instances of the same (or similar) ligands in different poses. 3DLigandSite originally predicted any residue within 0.8 A of at least 25 % of the ligands in a cluster to be part of the binding site. This has been replaced by the introduction of a logistic regression classifier (detailed below) to perform this final prediction step. This also associates a confidence score (range 0-1) with each residue in the predicted binding site.

Generation of the library of ligand-bound protein structures

To generate the library of biologically relevant protein binding sites, protein structures were extracted from the PDB and filtered to retain only those containing ligands classed as cognate by FireDB [23]. The library focusses on monomeric proteins. Where binding sites were located in the interface between two proteins the multimer was split into monomeric structures and the ligands associated with both of the monomers. The protein structures were clustered, and the ligands from proteins in each cluster mapped onto a representative structure to reduce search time. The amino acid sequences of the retained structures were clustered using CD-HIT [48] using an 80 % sequence identity threshold. The protein models in each cluster were then aligned to the cluster representative (obtained from CD-HIT) using TM-align [47], the ligands were superimposed onto the representative structure and

retained. An HHSearch [49] sequence database was built from the representative sequences for searching user-submitted protein sequences against.

Calculating residue conservation—To calculate residue conservation, HHBlits [50] was used to search the query sequence against the UniClust30 database [51]. The multiple sequence alignment was then used to calculate the Jensen-Shannon Divergence [52] conservation score.

Machine learning-based prediction of binding site residues—The machine learning step was introduced to predict accurately which residues are most likely to be part of the binding site around a cluster of ligands. An equal number of binding and non-binding residues on the query protein were used for training and testing. For each of these residues, a set of features was extracted and converted to a 0-1 range (Supplementary Table 1). Several features were considered for best determining binding propensity. The features included distance measurements to the ligand cluster, residue conservation and amino acid properties such as charge, hydrophobicity and van der Waals volume (Supplementary Table 1). Solvent accessibility scores were obtained from ProAct2 [53]. Distance-based features were calculated including the minimum, maximum, and average distance of each residue to ligands in the cluster, and the percentage of ligands in the cluster within 0.8 Å + Van der Waals radii of the amino acid.

Univariate feature selection was used to identify ligand contacts. The three distance features, and residue conservation were the most informative features for predicting ligand binding, as well as the negative charge residue feature for metal binding sites. A single distance metric was selected to avoid overtraining on a similar feature, resulting in the ligand contacts, minimum ligand distance, negatively charged, and residue conservation as the selected features.

The scikit-learn Python package was used to train support vector machines (SVMs) (Cortes and Vapnik, 1995), Extra-Trees, logistic regression, and random forest classifiers. The data were then fitted with optimum parameters from 100 random iterations and three cross-validation steps using GridSearchCV within scikit-learn. A randomly generated 80:20 train-test split was used to fit the models.

The training and test sets comprise monomers with cognate ligands bound. These structures were identified by filtering the PDB, clustering their sequences using MMseqs2 [54]) at a maximum sequence identity of 40 %. This resulted in 5223 metal and 4995 non-metal binding sites. A subset of 1600 metal and 1573 non-metal binding sites were randomly selected for testing and training. The remaining binding sites were used as a validation set to evaluate performance on the trained classifiers (that had not been used in training) (Supplementary Figure 1). The PDB identifiers and chains of all sequences used are provided in Supplementary Table 2.

Binding residues were classed as all residues within VDW radii + 0.8 Å of the ligand present in the protein structure, with all other residues classed as non-binding. This resulted in 1976 and 6950 metal and non-metal binding residues, respectively, and an equal number of

Europe PMC Funders Author Manuscripts

randomly selected non-binding residues were also randomly extracted (Supplementary Table 3), providing the positive and negative examples required for training the machine learning classifiers.

Evaluating 3DLigandSite Performance

The performance of 3DLigandSite was assessed using the validation set (see methods section), which contained 59203 and 16166 (Supplementary Table 3) non-metal and metal-binding residues, respectively, that had not been used in the testing or training of the classifiers. Performance was assessed using multiple measures of precision, sensitivity (recall), and the Receiver Operator Characteristic (ROC).

The logistic regression classifier performed best on the non-metal binding sites, with a an area under the receiver operating characteristic curve (AUROC) of 0.99, though a similar performance was observed for Extra-Trees and random forest classifiers (Figure 2A, Supplementary Table 4). For metal ligands, the logistic regression classifier performed best with an AUROC of 0.99 (Figure 2C, Supplementary Table 4). As the data set has a skewed distribution with many more negative examples than positive examples (i.e. many non-binding residues compared to those that are binding residues in each protein – Supplementary table 3), precision-recall metrics provide a better indication of performance [55,56]. 3DLigandSite obtained 92 % recall at 75 % precision for non-metal binding sites (Figure 2B) and 52 % recall at 75 % precision for metal-binding sites (Figure 2D). We compared the performance of the new version of 3DLigandSite with the original version [25]. The original 3DLigandSite did not make separate predictions for metal and non-metal ligands, so we assessed performance on the combined metal and non-metal binding sites. On the validation set the original 3DLigandSite obtained recall of 56% at 59% precision.

The performance of 3DLigandSite was also evaluated on the 70 targets used for assessment of binding site prediction in CASP8 [26], CASP9 [57] and CASP10 [58]. Using the sequence-based homology search 3DLigandSite obtained recall of 85% at 65% precision, and a Matthews' Correlation Coefficient (MCC; [59]) of 0.73. Performance using the structural-search option was comparable, with slightly lower recall of 80% at 67% precision and an MCC of 0.72. Structural search results at a range of TM-Score thresholds for inclusion of template structures are shown in Supplementary Table 5. On this dataset, the sequence-based search was not inferior to the structure-based search, although recent studies have suggested that structural searches are better at identifying related protein structures [60,61]. Given, the extra time taken to perform the structural search (approximately 4 hours per submission), the sequence search is recommended and is the default for the web server.

The 3DLigandSite Web Server

The 3DLigandSite web server is available at https://www.wass-michaelislab.org/3dligandsite. The web server is free to all without a login requirement. Users can select to submit either a protein sequence (in FASTA format) or a protein structure (in PDB format). Where a sequence is submitted, the first step of the prediction process is to obtain a model of the protein structure. To do this the PDB is first searched for a matching

structure, followed by AlphaFold DB [42]. Where a suitable model is not available in either database, Phyre2 [44] is used to generate a template-based model of the structure. The runtime for submissions that require Phyre2 is longer as modelling the protein structure is time-consuming, typically taking a few hours to complete. Where users submit a protein structure, the runtime is typically less than five minutes using default settings. Users who provide an email address receive an email upon submission and once their results are ready for viewing. The web server includes a help section that includes recordings that work users through both the submission process and interpretation of data in the results pages.

Results Output

3DLigandSite results pages are split into three main sections. Results are initially presented as a sequence view (Figure 3), which shows the amino acid sequence of the submitted protein, residue conservation, and a row for each cluster of ligands that has been identified as a potential binding site (Figure 3). This provides users with an easily interpretable view of the predicted binding sites.

The second section of results shows the cluster table, which includes details of the clusters identified, the number of ligands present in each cluster, and the number of structures that these ligands originate from. The ligands are represented by the three-letter codes from the mmCIF dictionary and are linked to the small molecule details in the PDB (Figure 3). Clusters are sorted according to the number of ligands present in the cluster. There is greater confidence that a cluster represents a binding site when there is evidence for this from multiple protein structures. The second table in this section contains a tab for each ligand cluster and lists the residues predicted to be in the binding site along with the conservation score, solvent accessibility, and the probability calculated by the logistic regression classifier.

The final section of the results page contains a Mol* molecular viewer (www.molstar.org) [62] that by default displays the protein structure in a cartoon format along with the ligands in the top-ranked cluster, highlighting the predicted binding site residues in red (Figure 3). The Mol* viewer enables users to inspect the predicted binding sites within the protein structure and offers multiple features for exploring the structure. The 3DLigandSite control panel to the right of the viewer provides easy to use functions such as changing the colour or format of the display of the ligands and the protein structure. Further functionality is available via the Mol* built-in options shown on the top right of the viewer. The control panel also includes a button enabling users to generate publication-quality images of the current display in the viewer.

Use Cases

As set out in the introduction, 3DLigandSite predictions have been widely used for a range of different biological and biomedical purposes. For example, with widespread use of sequencing technologies, there is extensive interest in the analysis of non-synonymous single nucleotide variants (nsSNVs). The aim here is to identify those nsSNVs that may alter protein structure and function and be associated with a phenotype such as a disease. Thus,

3DLigandSite has been used to analyse such nsSNVs for a range of diseases, from liver disease [31] to cardiomyopathies [33].

One application has been to study nsSNVs present in individuals with cystinuria, which is caused by variants in two genes, SLC7A9 and SLC3A1, that encode a dimeric amino acid transporter [63]. Cystinuria is caused by variants that affect the ability of this transporter to transport cystine into cells, which results in the formation of kidney stones. In a recent study [34, 64], 3DLigandSite was used to model the structure and ligand binding sites of the two encoded proteins and to analyse how the set of nsSNVs observed in a cohort of patients may affect transporter function and be linked with the severity of the disease that patients experienced. Figure 3 shows the protein b(0+)AT, which is encoded by SLC7A9, and the predicted amino acid binding sites in the protein.

Concluding Remarks

The 3DLigandSite web server provides free access to an easy-to-use resource for modelling small molecule binding sites in proteins. This widely used resource has been extensively updated to offer improved functionality and to reduce the run time of user submissions. Our benchmarking demonstrates that 3DLigandSite can obtain high recall with high precision, therefore accurately predicting binding sites in proteins that users are researching.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding

JEM was supported by an Engineering and Physical Sciences Research Council PhD Studentship. This research was funded in in part by the Wellcome Trust grants WT104955MA and 218242/Z/19/Z to MJES and by the BBSRC grant BB/J0105451/1 to MJES. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Data Availabilty

All data are provided in the manuscript or supplementary material.

References

1. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021; 49: D480–D489. DOI: 10.1093/nar/gkaa1100 [PubMed: 33237286]
2. Zhou N, Jiang Y, Bergquist TR, Lee AJ, Kacsoh BZ, Crocker AW, Lewis KA, Georghiou G, Nguyen HN, Hamid MdN, et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. Genome Biology. 2019; 20: 244–23. DOI: 10.1186/s13059-019-1835-8 [PubMed: 31744546]
3. Jiang Y, Oron TR, Clark WT, Bankapur AR, D'Andrea D, Lepore R, Funk CS, Kahanda I, Verspoor KM, Ben-Hur A, et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biol. 2016; 17: 184. doi: 10.1186/s13059-016-1037-6 [PubMed: 27604469]
4. Mukhopadhyay A, Borkakoti N, Pravda L, Tyzack JD, Thornton JM, Velankar S. Finding enzyme cofactors in Protein Data Bank. Bioinformatics. 2019; 35: 3510–3511. DOI: 10.1093/bioinformatics/btz115 [PubMed: 30759194]

5. Torrance JW, Macarthur MW, Thornton JM. Evolution of binding sites for zinc and calcium ions playing structural roles. Proteins. 2008; 71: 813–830. DOI: 10.1002/prot.21741 [PubMed: 18004751]

6. Zhao J, Cao Y, Zhang L. Exploring the computational methods for protein-ligand binding site prediction. Comput Struct Biotechnol J. 2020; 18: 417–426. DOI: 10.1016/j.csbj.2020.02.008 [PubMed: 32140203]

7. Capra JA, Singh M. Characterization and prediction of residues determining protein functional. Bioinformatics. 2008; 24: 1473–1480. DOI: 10.1093/bioinformatics/btn214 [PubMed: 18450811]

8. Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLOS Comput Biol. 2009; 5 e10000585 doi: 10.1371/journal.pcbi.1000585 [PubMed: 19997483]

9. Krivak R, Hoksza D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. J Cheminform. 2018; 10: 1–12. DOI: 10.1186/s13321-018-0285-8 [PubMed: 29340790]

10. Jendele L, Krivak R, Skoda P, Novotny M, Hoksza D. PrankWeb: a web server for ligand binding site prediction and visualization. Nucleic Acids Res. 2019; 47: W345–349. DOI: 10.1093/nar/gkz424 [PubMed: 31114880]

11. Santana CA, Silveira Sde A, Moraes JPA, Izidoro SC, de Melo-Minardi RC, Ribeiro AJM, Tyzack JD, Borkakoti N, Thornton JM. GRaSP: a graph-based residue neighborhood strategy to predict binding sites. Bioinformatics. 2020; 36: i726–i734. DOI: 10.1093/bioinformatics/btaa805 [PubMed: 33381849]

12. Jimenez J, Doerr S, Martínez-Rosell G, Rose AS, De Fabritiis G. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. Bioinformatics. 2017; 33: 3036–3042. DOI: 10.1093/bioinformatics/btx350 [PubMed: 28575181]

13. Aggarwal R, Gupta A, Chelur V, Jawahar CV, Priyakumar UD. DeepPocket: Ligand Binding Site Detection and Segmentation using 3D Convolutional Neural Networks. J Chem Inf Model. 2021; doi: 10.1021/acs.jcim.1c00799 [PubMed: 34374539]

14. Stepniewska-Dziubinska M, Zielenkiewicz P, Siedlecki P. Improving detection of protein-ligand binding sites with 3D segmentation. Sci Rep. 2020; 1: 5035. doi: 10.1038/s41598-020-61860-z [PubMed: 32193447]

15. Kandel J, Tayara H, Chong KT. PUResNet: prediction of protein-ligand binding sites using deep residual neural network. J Cheminform. 2021; 13: 65. doi: 10.1186/s13321-021-00547-7 [PubMed: 34496970]

16. Mylonas SK, Axenopoulos A, Daras P. DeepSurf: a surface-based deep learning approach for the prediction of ligand binding sites on proteins. Bioinformatics. 2021; 37: 1681–1690. DOI: 10.1093/bioinformatics/btab009 [PubMed: 33471069]

17. Lopez G, Maietta P, Rodriguez JM, Valencia A, Tress ML. firestar--advances in the prediction of functionally important residues. Nucleic Acids Res. 2011; 39: W235–41. DOI: 10.1093/nar/gkr437 [PubMed: 21672959]

18. Brylinski M, Skolnick J. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. Proc Natl Acad Sci USA. 2008; 105: 129–134. DOI: 10.1073/pnas.0707684105 [PubMed: 18165317]

19. Feinstein WP, Brylinski M. Enhanced Fingerprint-Based Virtual Screening Against Predicted Ligand Binding Sites in Protein Models. Molecular Informatics. 2014; 33: 135–150. DOI: 10.1002/minf.201300143 [PubMed: 27485570]

20. Wu Q, Peng Z, Zhang Y, Yang J. COACH-D: improved protein-ligand binding sites prediction with refined ligand-binding poses through molecular docking. Nucleic Acids Res. 2018; 46: W438–W442. DOI: 10.1093/nar/gky439 [PubMed: 29846643]

21. Roche DB, Buenavista MT, McGuffin LJ. FunFOLD2 server for the prediction of protein-ligand interactions. Nucleic Acids Res. 2013; 41: W3030–W307. DOI: 10.1093/nar/gkt498 [PubMed: 23761453]

22. Armstrong DR, Berrisford JM, Conroy MJ, Gutmanas A, Anyango S, Choudhary P, Clark AR, Dana JM, Deshpande M, Dunlop R, et al. PDBe: improved findability of macromolecular structure

data in the PDB. Nucleic Acids Res. 2020; 48: D335–D343. DOI: 10.1093/nar/gkz990 [PubMed: 31691821]

23. Maietta P, Lopez G, Carro A, Pingilley BJ, Leon LG, Valencia A, Tress ML. FireDB: a compendium of biological and pharmacologically relevant ligands. Nucleic Acids Res. 2013; 42: D267–72. DOI: 10.1093/nar/gkt1127 [PubMed: 24243844]
24. Ribeiro AJM, Holliday GL, Furnham N, Tyzack JD, Ferris K, Thornton JM. Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. Nucleic Acids Res. 2018; 46: D618–D623. DOI: 10.1093/nar/gkx1012 [PubMed: 29106569]
25. Wass MN, Kelley LA, Sternberg MJE. 3DLigandSite: predicting ligand-binding sites using similar structures. Nucleic Acids Res. 2010; 38: W469–73. DOI: 10.1093/nar/gkq406 [PubMed: 20513649]
26. Lopez G, Ezkurdia I, Tress ML. Assessment of ligand binding residue predictions in CASP8. Proteins. 2009; 77 (Suppl9) 138–46. DOI: 10.1002/prot.22557 [PubMed: 19714771]
27. Wass MN, Sternberg MJE. Prediction of ligand binding sites using homologous structures and conservation at CASP8. Proteins. 2009; 77 (Suppl 9) 147–151. DOI: 10.1002/prot.22513 [PubMed: 19626715]
28. Antczak M, Michaelis M, Wass MN. Environmental conditions shape the nature of a minimal bacterial genome. Nat Commun. 2019; 10: 3100. doi: 10.1038/s41467-019-10837-2 [PubMed: 31308405]
29. Nishiyama T, Sakayama H, de Vries J, Buschmann H, Saint-Marcoux D, Ullrich KK, Haas FB, Vanderstraeten L, Becker D, Lang D, et al. The Chara Genome: Secondary Complexity and Implications for Plant Terrestrialization. Cell. 2018; 74: 448–464. e24 doi: 10.1016/j.cell.2018.06.033 [PubMed: 30007417]
30. Kuhlmann FM, Robinson JI, Bluemling GR, Ronet C, Fasel N, Beverley SM. Antiviral screening identifies adenosine analogs targeting the endogenous dsRNA Leishmania RNA virus 1 (LRV1) pathogenicity factor. Proc Natl Acad Sci. 2017; 114: E811–E819. DOI: 10.1073/pnas.1619114114 [PubMed: 28096399]
31. Chambers JC, Zhang W, Sehmi J, Li X, Wass MN, Van der Harst P, Holm H, Sanna S, Kavousi M, Baumeister SE, et al. Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. Nat Genet. 2011; 43: 1131–1138. DOI: 10.1038/ng.970 [PubMed: 22001757]
32. Bernkopf M, Webersinke G, Tongsook C, Koyani CN, Rafiq MA, Ayaz M, Muller D, Enzinger C, Aslam M, Naeem F, et al. Disruption of the methyltransferase-like 23 gene METTL23 causes mild autosomal recessive intellectual disability. Hum Mol Genet. 2014; 23: 4015–4023. DOI: 10.1093/hmg/ddu115 [PubMed: 24626631]
33. O'Grady GL, Best HA, Sztal TE, Schartner V, Sanjuan-Vazquez M, Donkervoort S, Neto OA, Sutton RB, Ilkovski B, Romero NB, et al. Variants in the Oxidoreductase PYROXD1 Cause Early-Onset Myopathy with Internalized Nuclei and Myofibrillar Disorganization. The American Journal of Human Genetics. 2016; 99: 1086–1105. DOI: 10.1016/j.ajhg.2016.09.005 [PubMed: 27745833]
34. Martell HJ, Wong KA, Martin JF, Kassam Z, Thomas K, Wass MN. Associating mutations causing cystinuria with disease severity with the aim of providing precision medicine. BMC Genomics. 2016; 18: 550. doi: 10.1186/s12864-017-3913-1 [PubMed: 28812535]
35. Papalardo M, Wass MN. VarMod: modelling the functional effects of non-synonymous variants. Nucleic Acids Res. 2014; 42: W331–6. DOI: 10.1093/nar/gku483 [PubMed: 24906884]
36. Ho CH, Frommer WB. Fluorescent sensors for activity and regulation of the nitrate transceptor CHL1/NRT1.1 and oligopeptide transporters. eLife. 2014; 3 e01917 doi: 10.7554/eLife.01917 [PubMed: 24623305]
37. Bojkova D, McGreig JE, McLaughlin KM, Masterson SG, Antczak M, Widera M, Krahling V, Ciesek S, Wass MN, Michaleis M, Cinatl J Jr. Differentially conserved amino acid positions may reflects differences in SAR-CoV-2 and SARS-CoV behaviour. Bioinformatics. 2021; 37: 2282–8. DOI: 10.1093/bioinformatics/btab094 [PubMed: 33560365]
38. Agrawal A, Varshney R, Pathak M, Patel SK, Rai V, Sulabh S, Gupta R, Solanki KS, Varshney R, Nmmanapalli R. Exploration of antigenic determinants in spike glycoprotein of SARS-CoV2

and identification of five salient potential epitopes. Virusdisease. 2021; 32: 1–10. DOI: 10.1007/s13337-021-00737-9 [PubMed: 33644261]

39. Venkateshan M, Muthu M, Suresh J, Kumar RR. Azafluorene derivatives as inhibitors of SARS CoV-2 RdRp: Synthesis, physicochemical, quantum chemical, modeling and molecular docking analysis. J Mol Struct. 2020; 1220 128741 doi: 10.1016/j.molstruc.2020.128741 [PubMed: 32834110]
40. PDBe-KB consortium. PDBe-KB: a community-driven resource for structural and functional annotations. Nucleic Acids Res. 2020; 48: D344–D353. DOI: 10.1093/nar/gkz853 [PubMed: 31584092]
41. PDBe-KB consortium. PDBe-KB: collaboratively defining the biological context of structural data'. Nucleic Acids Research. 2022; 50: D534–D542. DOI: 10.1093/nar/gkab988 [PubMed: 34755867]
42. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, Zidek A, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Rese. 2022; 50: D439–D444. DOI: 10.1093/nar/gkab1061 [PubMed: 34791371]
43. Jumper J, Evans E, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Zidek A, Potapenko A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021; 596: 583–89. DOI: 10.1038/s41586-021-03819-2 [PubMed: 34265844]
44. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015; 10: 845–858. DOI: 10.1038/nprot.2015.053 [PubMed: 25950237]
45. Ortiz AR, Strauss CEM, Olmea O. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci. 2002; Nov. 2606–21. doi: 10.1110/ps.0215902 [PubMed: 12381844]
46. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics. 2005; 21: 951–960. DOI: 10.1093/bioinformatics/bti446 [PubMed: 15531603]
47. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005; 33: 2302–2309. DOI: 10.1093/nar/gki524 [PubMed: 15849316]
48. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012; 28: 3150–3152. DOI: 10.1093/bioinformatics/bts565 [PubMed: 23060610]
49. Steinegger M, Meier M, Mirdita M, Vohringer H, Haunsberger SJ, Soding J. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics. 2019; 20: 473–15. doi: 10.1186/s12859-019-3019-7 [PubMed: 31521110]
50. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 2012; 9: 173–175. DOI: 10.1038/nmeth.1818 [PubMed: 22198341]
51. Mirdita M, den Driesch von L, Galiez C, Martin MJ, Soding J, Steinegger M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res. 2017; 45: D170–D176. DOI: 10.1093/nar/gkw1081 [PubMed: 27899574]
52. Capra JA, Singh M. Predicting functionally important residues from sequence conservation. Bioinformatics. 2007; 23: 1875–82. DOI: 10.1093/bioinformatics/btm270 [PubMed: 17519246]
53. Williams MA, Goodfellow JM, Thornton JM. Buried waters and internal cavities in monomeric proteins. Protein Science. 1994; 3: 1224–1235. DOI: 10.1002/pro.5560030808 [PubMed: 7987217]
54. Steinegger M, Soding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature Biotechnology. 2017; 35: 1026–8. DOI: 10.1038/nbt.3988 [PubMed: 29035372]
55. Wass MN, Sternberg MJE. ConFunc - functional anotation in the twilight zone. Bioinformatics. 2008; 24: 798–806. doi: 10.1093/bioinformatics/btn037 [PubMed: 18263643]
56. Davis, J; Goadrich, M. The relationship between Precision-Recall and ROC Curves; 23rd International Conference on Machine Learning (ICML); Pittsburgh, PA, USA. June 26–28, 2006;;

57. Schmidt T, Haas J, Cassarino TG, Schwede T. Assessment of ligand-binding residue predictions in CASP9. 2011; 79 (Suppl 10) 126–36. DOI: 10.1002/prot.23174 [PubMed: 21987472]
58. Cassarino TG, Bordoli L, Schwede T. Assessment of ligand binding site predictions in CASP10. Proteins. 2014; 82 (Suppl 2) 154–63. DOI: 10.1002/prot.24495 [PubMed: 24339001]
59. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta. 1975; 405: 442–51. DOI: 10.1016/0005-2795(75)90109-9 [PubMed: 1180967]
60. Chen J, Guo M, Wang X, Liu B. A comprehensive review and comparison of different computational methods for protein remote homology detection'. Briefings in Bioinformatics. 2018; 19: 231–244. doi: 10.1093/bib/bbw108 [PubMed: 27881430]
61. Yan R, Xu D, Yang J, Walker S, Zhang Y. A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. Sci Rep. 2013; 3: 2619. doi: 10.1038/srep02619 [PubMed: 24018415]
62. Sehnal, D; Rose, A; Koca, J; Burley, S; Velankar, S. Mol*: Towards a common library and tools for web molecular graphics; Workshop on Molecular Graphics and Visual Analysis of Molecular Data; 2018.
63. Thomas K, Wong K, Withington J, Bultitude M, Doherty A. Cystinuria-a urologist's perspective. Nat Rev Urol. 2014; 11: 270–277. DOI: 10.1038/nrurol.2014.51 [PubMed: 24662732]
64. Wong KA, Wass M, Thomas K. The Role of Protein Modelling in Predicting the Disease Severity of Cystinuria. Eur Urol. 2016; 69: 543–4. DOI: 10.1016/j.eururo.2015.10.039 [PubMed: 26589650]

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

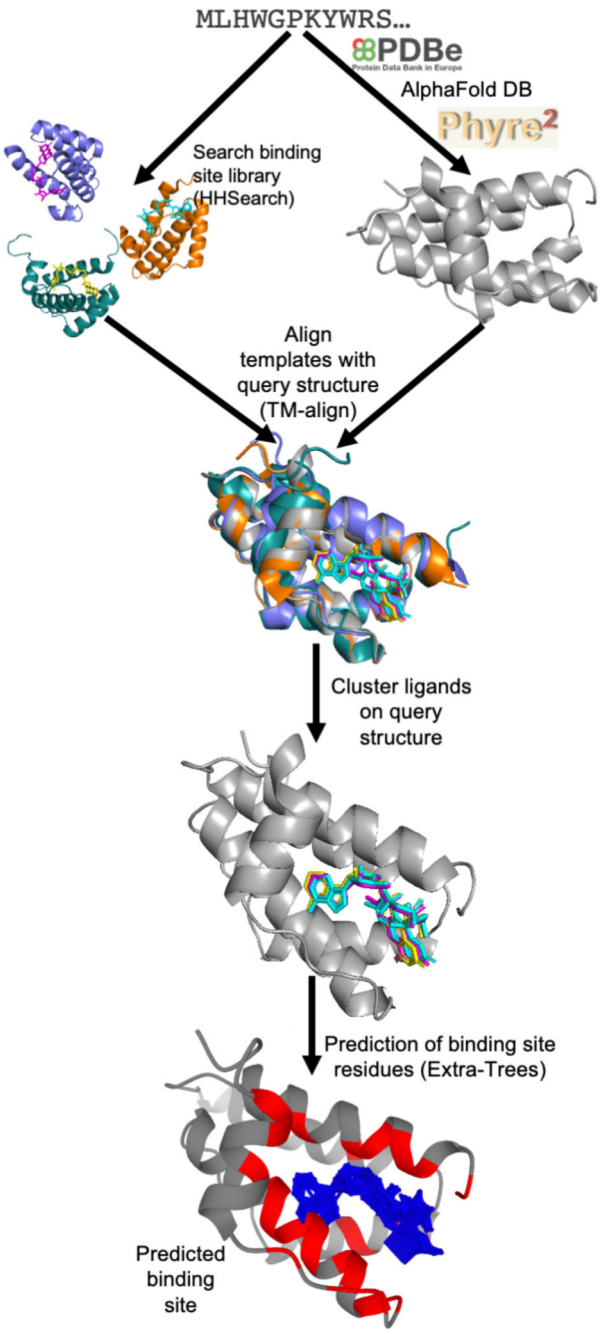

Figure 1. An overview of the 3DLigandSite method.

Users submit either a protein sequence or structure. Where sequences are submitted the PDBe and AlphaFold DB are searched for a matching structure, where one is not available Phyre2 is used to model the 3D structure. HHsearch is used to search a sequence library of protein structures with ligands bound. Hits from this search are aligned with the structure of query protein, the ligands from these structures are clustered. Each cluster of ligands represents a potential binding site in the query protein. A machine learning classifier is used to predict which of the residues around the cluster are likely to form part of a binding site.

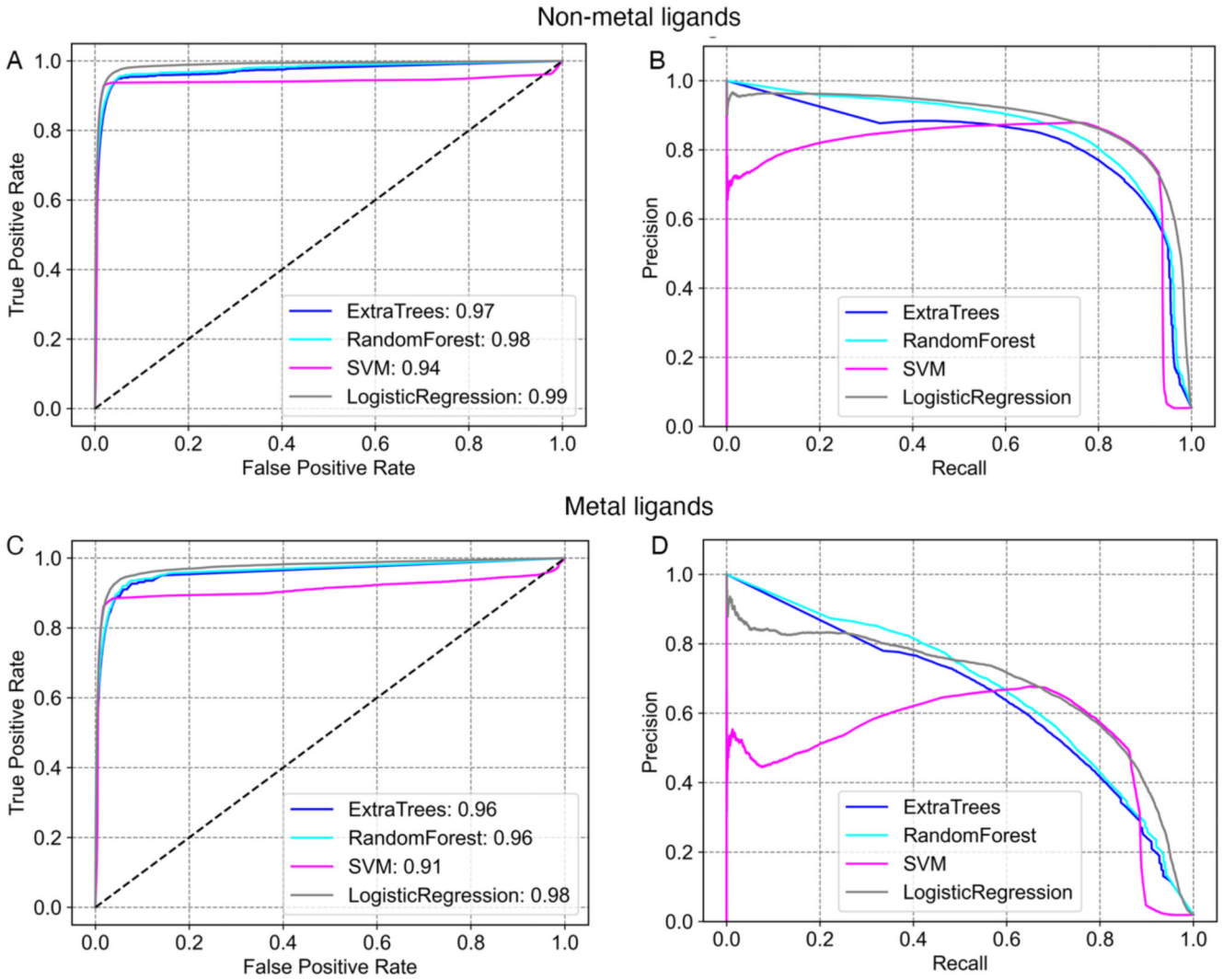

Figure 2. Benchmarking the 3DLigandSite machine learning classifier.

Receiver operator characteristic (ROC) curves and Precision-Recall curves are shown for the prediction of binding sites of non-metal (A and B) and metal (C and D) ligands.

Europe PMC Funders Author Manuscripts

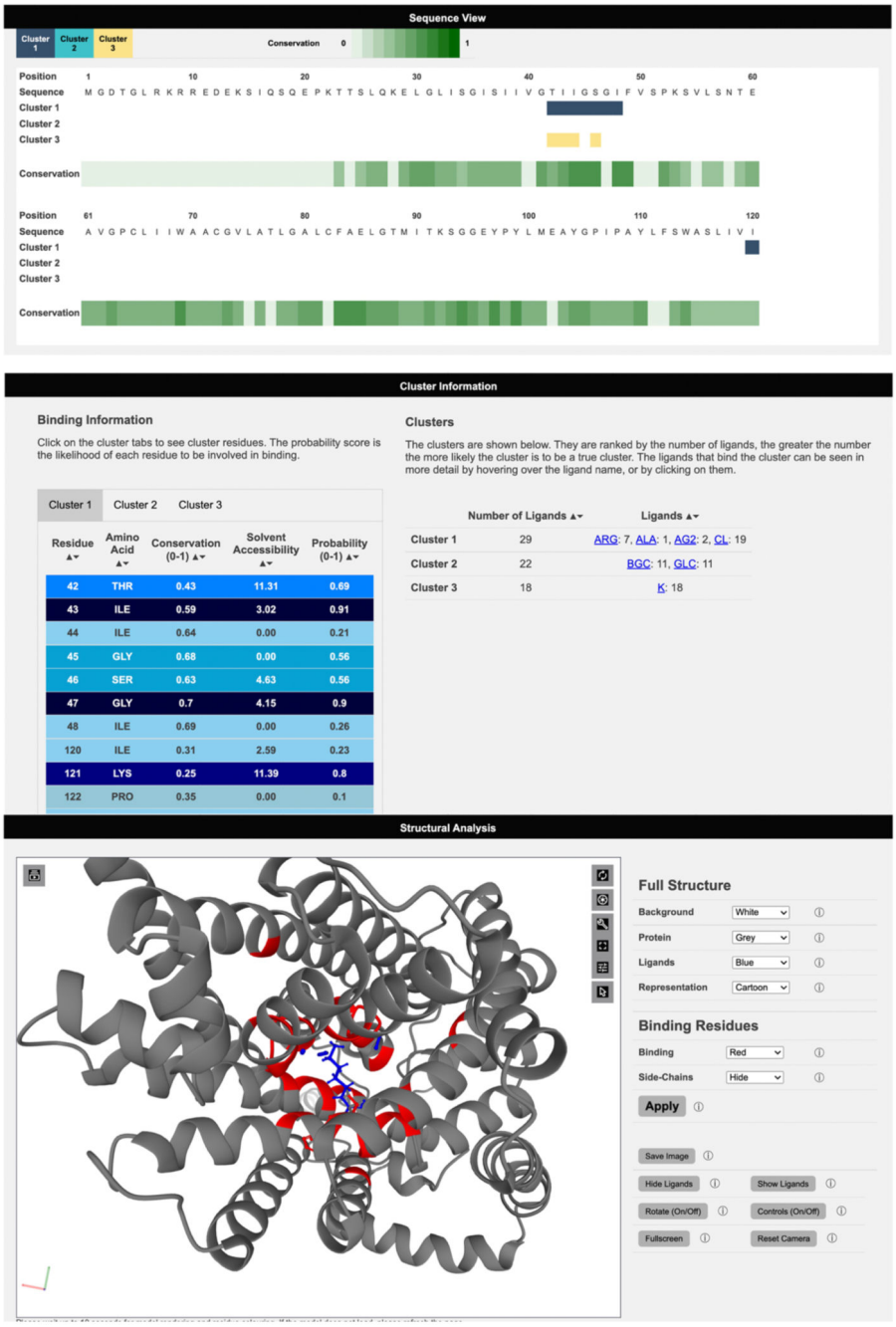

Figure 3. Viewing results on the 3DLigandSite web server.

Results are presented in 3 main sections. A sequence view, which maps sequence conservation and the different clusters identified onto the protein sequence. Secondly, details of the clusters, including the number of ligands and type of ligand are displayed as well as a table listing the residues predicted to form the binding site for each cluster. Finally, the structural analysis section includes a Mol* molecular viewer to visualise the protein, the

predicted binding site and the clusters used to make the predictions. A separate control panel (on the right) enables users to easily modify the display.

Europe PMC Funders Author Manuscripts