Computational Approaches to Predict Protein Functional Families and Functional Sites

Rauer C[1†], Sen N[1†], Waman VP[1†], Abbasian M[1], Orengo CA[1,*]

1 Institute of Structural and Molecular Biology, University College London, London, WC1E 6BT, UK
† These authors contributed equally
* To whom correspondence should be addressed: c.orengo@ucl.ac.uk

Abstract

Understanding the mechanisms of protein function is indispensable for many biological applications, like protein engineering and drug design. However, experimental annotations are sparse and therefore, theoretical strategies are needed to fill the gap. Here, we present the latest developments in building functional subclassifications of protein superfamilies, and using evolutionary conservation to detect functional determinants e.g. catalytic-, binding- and specificity determining residues important for delineating the functional families. We also briefly review other features exploited for functional site detection and new machine learning strategies for combining multiple features.

Introduction

Protein function space is poorly characterised. Less than 10% of proteins in UniProt are experimentally characterised and the characterisation of functional sites in proteins is even more sparse (<1%), see Figure 1a. The recent explosion of protein sequence data through metagenomics initiatives suggests that there is likely to be even more functional diversity than currently captured in UniProt. For example, exploration of the alpha/beta hydrolases showed a 5-fold expansion in functional families identified in bacterial communities in oceans and wastewater[1]. Computational approaches are much needed to survey and characterise this dark function space. In this review we consider recent strategies for predicting functionally similar proteins and for exploiting this information, together with other key data, to predict their functional sites.

Methods for classifying proteins into functional families

Several resources exist for classifying protein families. They either classify the entire protein (e.g. PANTHER[1], TigrFam[2] or focus on the single domain components (e.g. Pfam[3], SCOP[4], CATH[5])). However, most do not sub-classify by function. Here, we review recent developments in classification of protein functional families (funfams) and the exploitation of

[1] Unpublished data

family data to detect functional determinant residues (e.g. catalytic, binding, specificity determining residues).

Comparing the different approaches is complicated as most have only been tested on a few selected superfamilies, and there is no widely used benchmark. Experimental data from the Enzyme Resource[6], GO[7] and UniProt[8] can be used. A metric developed by Sjolander and co-workers[9] captures both family purity and whether families suffer from over-splitting. Further assessments of methods can be obtained if developers participate in CAFA (Critical Assessment of protein Function Annotation)[10–12], where the function of a set of experimentally uncharacterised proteins has to be predicted.

Many approaches for funfam generation build on early pioneering strategies[13–15] that derive a phylogenetic tree from the multiple sequence alignment (MSA) of the superfamily, and then split the tree to give functional subgroups. Other strategies use pairwise sequence comparisons to derive networks which are then divided into functional clusters (see Figure 1 (B)). In very large structurally diverse superfamilies generation and analysis of the MSA can be challenging. Recent analyses of the CATH[16] superfamily classification shows that whilst ~74% of superfamilies are small (<1000 sequences) and structurally conserved (see Figure 1(a)) for the largest 10%, there can be considerable structural variation. These superfamilies are highly populated and account for 50% of all CATH domains. Relatives can vary 3-fold or more in size and superpose with >9Å RMSD. These structural changes can introduce large indels into the MSA and making analysis of MSA problematic.

Amongst unsupervised methods deriving funfams from the MSA, Rivoire et al.[17] use statistical coupling analysis to determine amino acid coevolution, based on the MSA. Using spectral decomposition they can determine functional families in the correlation matrix. This approach has been extended by other groups[18–20]. On the other hand, the strategy of Neuwald et al.[21] copes with the MSA challenge by using Bayesian Partitioning with Pattern Selection (BPPS), to partition the input alignment into a hierarchically nested series of MSAs, based on correlated residue patterns that are distinctive for each subgroup. By exploiting structure data their Structurally Interacting Pattern Residues' Inferred Significance (SIPRIS) method uses 3D residue clusters for refining selection of functional subgroups.

To tackle the problem of large, structurally variable superfamilies, GeMMA[22] avoids MSAs and instead generates a tree from the leaves upward, by iteratively comparing Hidden Markov models (HMMs) derived from sequence clusters in the superfamily, and merging the most similar clusters after each iteration, followed by regeneration of the MSA and HMM. FunFamer[23] then cuts the tree to give CATH-FunFams by exploiting GroupSim[24] to find function determining residues differentially conserved between FunFams (see Figure 1 (c)). Benchmarking showed these residues were highly enriched in experimentally known

functional sites[25]. FunFams were also validated using experimental data from Enzyme Commission (EC) and Structure Function Linkage Database (SFLD).

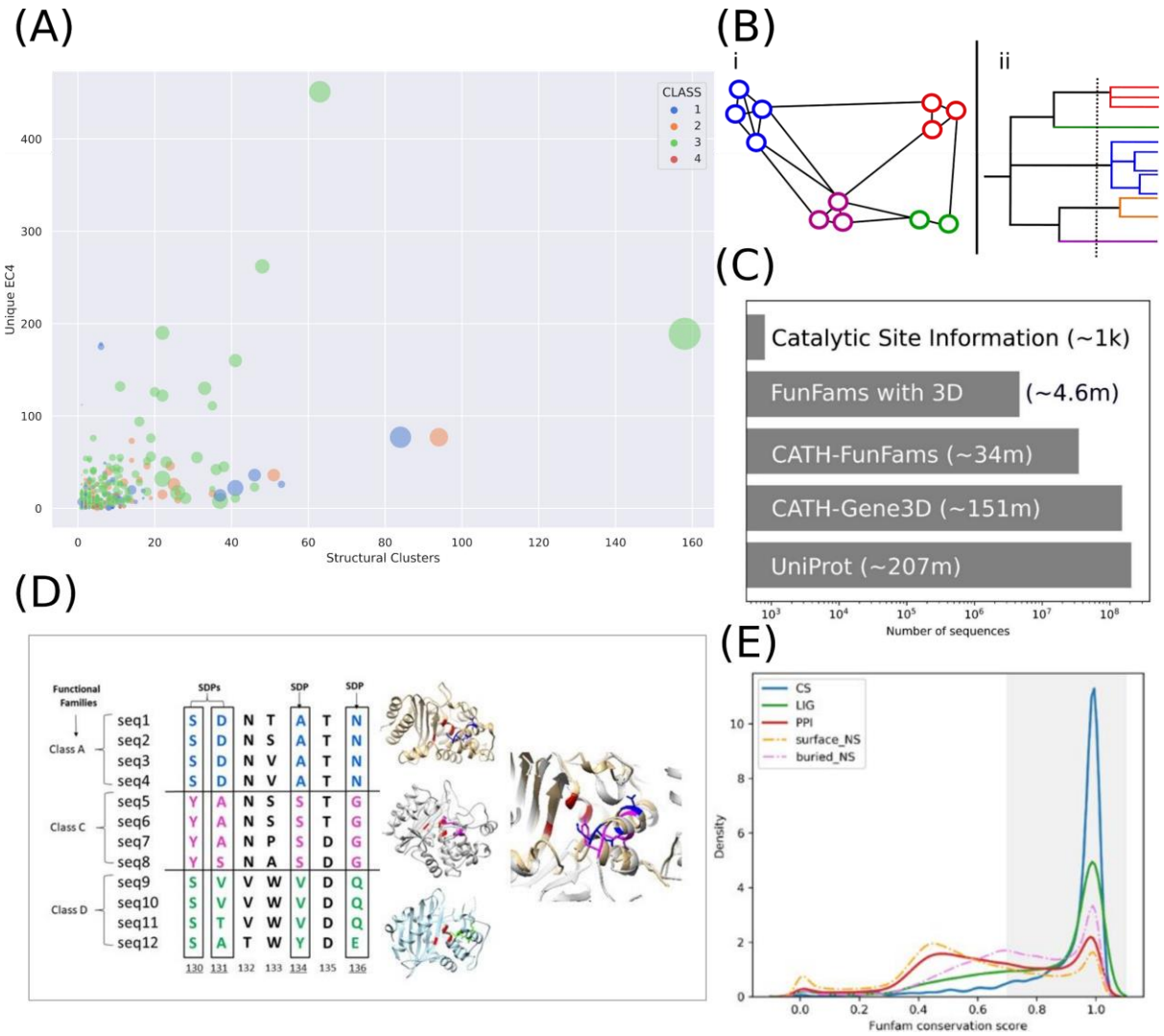

Figure 1 - (A) Scatter plot over CATH superfamilies showing the number of structural clusters vs number of unique EC4 terms in that family. The dot size indicates the number of sequences in the superfamily, and the color shows which class it belongs to. (B) Schematic description of the two main uses of treating the vast amount of sequence data for functional subclassification: (i) Sequence similarity network and (ii) Phylogenetic tree (C) The statistics on number of sequences in UniProt (at time of submission), in v4.3 of CAH-Gene3D, in CATH Functional families in functional families with a 3d representative and with known catalytic sites in the catalytic site atlas[26]. (D) Illustration of functional sub-classification and a selected set of sub-family specific Specificity Determining Position (SDPs), in beta-lactamase superfamily. The beta-lactamase superfamily comprises 3 well-known functional classes : Class A , C and D. The selected set of SDPs for each Class are colored in blue, magenta and green, respectively. These are mapped on the corresponding class's representative structure (Class A: 1shvA, Class C: 1zkjA, Class D: 1m6kA). The catalytic sites are depicted in red. The example is illustrated using results based on Lee et al.[27] (E) Distribution of sequence conservation scores calculated using Scorecons[28] (value between 0 and 1) for catalytic site (CS), ligand-

binding site (LIG) and PPI site residues within CATH FunFams. These are compared to buried non-site residues (buried_NS) and surface residues (surface_NS). Residues with scorecons[28] value ≥ 0.7 are considered to be conserved (grey region in the figure) [duplicated with permission][29]

Rather than analysing the sequence-based trees derived from MSAs, some approaches build a sequence similarity network (typically from pairwise BLAST comparisons), which is then broken into functional modules[30,31] (Figure 1(C)). Clusters can be refined by focussing on key functional regions in the protein. For example, TuLIP (Two-Level Iterative clustering Process)[32], generates an active site profile, comprising residues within 10 Å of the protein active site. Profiles are built for subgroups with different active site characteristics. While this limits the method to regions of protein family space where active site information exists, it drastically reduces the noise compared to using the whole sequence for comparison. Accuracy is improved in their more recent MISST (Multi-level Iterative Sequence Searching Technique) method[33], which scans the active site profile against GenBank to extend the data and thereby refine the clusters. MISST splits the TULIP-generated funfams if two distinct families emerge within the active site profile as more sequences are added. Other approaches exploiting structure have used genetic algorithms[34] to explore different combinations of superfamily descriptors, ranging from the whole structure to active site residues.

Since domains typically occur in diverse multi-domain contexts that can modify functional properties[35] a rather unique approach exploits protein interaction networks and domain-mediated interactions to detect functional subgroups or funfams[36]. However, this is clearly limited for single-domain superfamilies, which constitute ~10% of CATH superfamilies and ~40% of prokaryotic proteins[37]. Another interesting new approach (SignDy[38]), incorporates information on protein normal modes in the sub-clustering and achieved reasonable agreement with CATH-FunFams for a variety of superfamilies.

The recent vast expansion in sequence data from metagenome studies (with the MGnify resource[39] 20-fold larger than UniProt) has given an exciting stimulus to characterise families but the scale of the data severely complicates the use of MSAs. New supervised strategies, recognising relatives for previously characterised families, exploit multiple alignment free analyses and use deep learning to recognise family characteristics from large-scale sequence data. For example, DeepFam[40], which trains on existing protein family classifications, like Pfam, and DeepNog[41] which like DeepFam uses convolution neural networks to extract informative subsequence patterns from sets of proteins. Both methods use supervised strategies that don't rely on feature extraction, e.g. k-mer frequencies, but just use the raw protein sequences as input. Other recent deep learning strategies have exploited ProtBert[42] to characterise CATH FunFams and then used the resulting embedding to generate a finer, more accurate, functional classification[43].

Methods detecting Specificity determining positions (SDPs)

SDPs have been categorised into 3 types[44] - (I) residues conserved in one subfamily but variable in other subfamilies, (II) residues differentially conserved between distinct subfamilies (see Figure 1(D)), (III) marginally conserved, where no apparent conservation of residues occurs but residues in subfamilies do have similar, distinct, physico-chemical/structural properties. There is considerable interest in these sites not only because mutations in them are frequently associated with human diseases[45] but also because mutations can be used to enhance efficacy e.g. in industrial enzymes[46].

Many functional sub-classification methods are guided by SDP detection (already described above) and so report these sites. Their power mainly comes from sequence data (up to 300-fold more than structure data, depending on the superfamily) and entropy/mutual information[44,47,48] to detect residues differentially conserved between subgroups. However, structural data can enhance performance by upweighting sites clustering in 3D (e.g. ETtrace[13], BPPS-SIPRIS[21], TULIP[32]).

Zebra3D[49] exploits 3D-alignments and cluster analysis to detect subfamily-specific patterns in the local structure i.e. residues/loops/secondary structure fragments. In addition to entropy-based scoring, information on residue physicochemical properties is used to enhance detection[47,50]. ASSP[27] exploits CATH-FunFams but generates a finer functional classification by sampling residue combinations within a 3D-sphere centred on the catalytic residues.

As mentioned already above for funfam identification, performance can be enhanced by using co-variation analyses to find residues forming distinct networks of correlated residues indicating structural features specific to a particular subfamily (see Table 1). In NetworkStats[51], MSAs are used to generate bipartite networks of proteins and residues and co-variation used to link residues and give an effective mapping of local conservation[52].

Whilst most approaches need to know functional sub-groupings to detect SDPs, DARC (Deep Analysis of Residue Constraints)[53] applies co-variation analyses to subgroups generated using the BPPS-SIPIRIS[20] unsupervised approach. SDPs were validated using experimental binding and hydrolysis assays. Sequence Re-weighting (SR)[54] also uses residue couplings and thereby revise weights associated with subfamilies.

Table 1: Recently developed computational methods for detection of SDP (specificity-determining Positions).

Name of method	**Website**	**Techniques used**	**Features used**				**Size of input sequence dataset**	**Speciality of the method**
			3D	Residue Covariation	Protein-protein interaction	Evolutionary/MSA		

NetworkStats[51]	http://www.biocomp.icb.ufmg.br/biocomp/software-and-databases/networkstats/ web-server: http://bioinfo.icb.ufmg.br/conan/	Graph theory		✓		User-defined / Hmmerscan against Pfam	Up to ~40,000	
BPPS-SIPRIS[21]	https://www.igs.umaryland.edu/labs/neuwald/software/sipris/	Statistical modelling	✓	✓		Generates very large MSA, using MAPGAPS search against databases e.g. NCBInr, env_nr	Very large (up to 500,000)	Detects interacting networks of residues, including those that are remote from protein active sites.
Deep Analysis of Residue Constraints (DARC)[53]	https://www.igs.umaryland.edu/labs/neuwald/software/darc/	Statistical modelling	✓	✓		Generates very large MSA, using MAPGAPS search against databases e.g. NCBInr, env_nr	Very large (up to 500,000)	Identifies 3D-clusters of putative SDPs
Sequence-Reweighting (SR)[54]	NA	Statistical modelling	✓	✓		User defined e.g. obtained from Pfam	Very large (up to 500,000)	Uses sequence reweighting when sampling
Zebra2[50]	https://biokinet.belozersky.msu.ru/zebra2	Statistical sampling	✓			User-defined / similarity search using Mustguseal against databases (e.g. UniProtKB / Swiss-Prot+ TrEMB), entropy analysis	Up to 15,000	
Xdet[55]	http://csbg.cnb.csic.es/pazos/Xdet/	Random Walk on protein networks and statistical sampling	✓		✓	Uses query sequence for BLAST search (E <0.01), entropy analysis	Large	Uses protein-protein interactome to compare protein partners
Group specific method[56]	http://naegle.wustl.edu/software	MSA sub-sampling				User-defined (e.g. Lac family curated using AlloRep database)	Small datasets	Group-conservation weighted SDP scoring
ASSP[27]	NA	Combinatorial residue sampling on structure	✓			User-defined (e.g. CATH functional families)	Large	Uses active site structure profile i.e. residues that are 8Å
Zebra3D[49]	https://biokinet.belozersky.msu.ru/Zebra3D	3D-based clustering	✓			User defined 3D-structural alignment or search against PDB	Small (Upto 128 structures)	Uses multiple 3D structure alignment of proteins

Prediction of Other Protein Functional Sites

Other important functional sites include catalytic and ligand binding which we briefly review below. Post-translational modification, allosteric and protein-protein interface sites, also very important, have been reviewed recently elsewhere [57–59].

As well as integrating computationally predicted sites, major resources like PDB-eKB[60] and UniProt-KB[8] provide data on experimental literature-curated sites that can be used for benchmarking and validation.

The sequence/structure features used in functional site detection are summarised in Table 2. For many methods using extensive homologue data (i.e. protein family, funfams) tends to enhance performance. Figure 1(E) illustrates that catalytic and binding residues have strong evolutionary signals compared to interfaces[29] and tools exploiting evolutionary data tend to

rank it as a major feature (Table 2). As with SDPs, a common technique to reduce false positives is to check if predicted sites cluster in 3D, as functional residues typically co-locate (Figure 2).

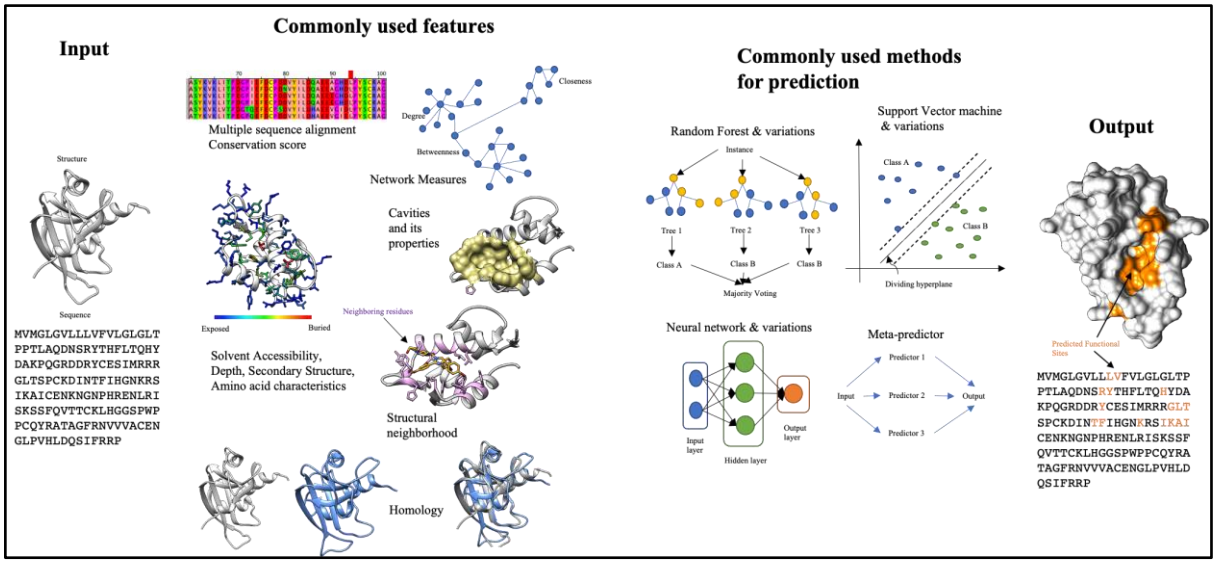

Figure 2 - Schematic showing the prediction of functional sites using some commonly used protein features and prediction methods. Most of these tools either rely on sequence/structural homology to a known template and/or machine learning (which depends upon various features to distinguish a functional site from others). The input for the functional site predictor (depending on the tool) can be in the form of a sequence/structure. The commonly used features for the prediction include sequence (multiple sequence alignment, evolutionary information, amino acid characteristics etc.), structural (solvent accessibility, depth, secondary structure, cavities, structural neighbourhood etc.) and/or network features of the protein structure (closeness, degree, betweenness etc.). Different combinations of these features and others are then used by various machine learning and deep learning techniques and their implementations such as random forest, support vector machine, neural network etc to train a prediction model. Some tools are meta-predictors which use predictions from multiple predictors to provide a final consensus output.

Increases in the underpinning data and advances in machine learning (ML), mean most approaches now exploit ML, but a significant challenge is the need to address the data imbalance between known functional residues and non-site residues. Some tools do this by oversampling binding sites and under-sampling non-binding sites[29] or using algorithms like AdaBoost[61,62]. Meta-predictors are frequently used to enhance performance by combining outputs from multiple tools to boost confidence where they agree (e.g. CSmetaPred_poc[63]). With the increased availability of data, deep learning techniques such as convoluted neural networks (CNN). graph convolution network (GCN) are increasingly being applied.

Catalytic sites:

Catalytic Site Atlas ((M-CSA[26]), reports predicted sites inherited from very close homologues by sequence searches. Structural information can enhance performance for distant homologues e.g. GASS-WEB[64] which uses a genetic algorithm to improve 3D-superposition of active sites, effectively exploring mutations between target and template.

Most other approaches exploit ML (see Table 2) to learn specific structure/sequence features (e.g. see Figure 2, Table 2). Information on 3D-environments e.g. contact networks (PREvaIL[65], FunSite[29]) is clearly helpful (though analyses showed evolutionary conservation was still a major feature). Usage of evolutionary coupling as in bindPredict[66] have shown to be useful. Significant improvements have been achieved recently using image processing techniques to represent the structure as 3D voxels. FSCNN[67] exploits such data using 3D Convolution Neural Networks (CNNs).

Ligand binding sites:

Metal, small-molecule, peptide and DNA/RNA binding sites are often valuable as drug targets. As well as predicted sites, known sites are provided by IBIS[68] and BioLip[69] which can be used for benchmarking. Many approaches exploit evolutionary conservation combined with other sequence/structural features (see Table 2). Below we highlight a few interesting new features and methodologies.

1. *Small molecule binding sites:*

Graph representation of residue neighbourhoods can be powerful, as graphs can be annotated with other residue features e.g. physico-chemical properties (GRaSP[70]). Most methods exploit ML, with random forest- and gradient boosted-decision trees being popular strategies (e.g. PrankWeb[71], Funsite[29], GRaSP[70], I-LBR[72]). FunSite captures conservation data using functional families (CATH-FunFams) as known binding sites are ~6 times more consistent between FunFam relatives than between general homologues[73].

With the increase in 3D data, more powerful deep learning models (e.g. CNNs) are being applied (eg DeepCSeqSite[74], DeepConv-DTI[75], DeepSite[76], DeepDrug3D[77], Kalasanty[78], DeepSurf[79]). Advances in image processing have been exploited by methods that represent the whole protein structure or surfaces/pockets as a 3D-image with voxels annotated with properties of the atoms (see Table 2). Additional features like conservation (DeepCSeqSite) and local residue contacts (DeepConv-DTI) can be included. DeepFRI[80], uses Graph Convolutional Networks (GCN) applied to residue contact maps (represented as graphs) and residue features learnt from 10 million PFam sequences (using a long short-term memory language model).

2. *Other sites:*

Recently developed predictors for other types of binding sites (peptide, metal, nucleic acid) are summarised in Table 2. Again, ML is widely used. Some novel features include the use of intrinsic disorder for peptide-binding sites (SVMPep[81]). Local neighbourhood features have been used to improve performance for metal-binding sites (mFASD[82]). For nucleic acid-binding, DNAPred uses a two-stage imbalance learning algorithm (Ensembled Hyperplane SVM), which takes care of the data imbalance between sites and non-sites. DRNApred[83], which predicts both DNA and RNA-binding sites, uses a two layer architecture to penalise cross prediction. The second layer re-predicts by checking the sequence neighbourhood. Another tool ProNA20[84] uses natural language processing along with other ML tools and homology to predict if a protein will bind to nucleic acid/protein and the binding site residues. In JET2DNA[85] , which does not use ML, local and global surface geometries are calculated by circular variance and feature score combinations are chosen depending on the conservation and curvature of the surface residues, making it more robust to stoichiometry and conformational changes.

Table 2: Computational methods for prediction of protein functional sites and allosteric sites. The table lists recently developed tools for the prediction of functional sites (catalytic and ligand binding sites (metal ion, small molecule, nucleic acid, peptide binding sites)). For each method, the table provides the technique and commonly used features used, any unique speciality and details on its usage of evolutionary information. [@]In the absence of webserver or stand-alone version, we write NA. [!]Solvent accessibility or secondary structure is predicted from sequence. [$]Involves a post processing step where the predicted residues are checked to see if they are structural neighbours, predicted residues without neighbours are removed from the final prediction. [*]Techniques that show evolutionary features are the top performing features or are important for the tool's prediction capabilities have been italicised. Default value for PSI-BLAST involves 3 iterations with an e-value cut-off of $1e^{-03}$.[#]The technique used by the tools to take care of data imbalance between the binding and non-binding sites have been mentioned in bold. Tools using image processing have been reported in italics.

Category of methods	**Name**	**Webserver/ Stand-alone version[@]**	**Description of techniques used**	**Input type**		**Common features used**							**Clustering of predicted residues on 3D structure[$]**	**Speciality of the method[#]**
						Residue features	Solvent Accessibility	Secondary structure	Pocket-based	3D neighborhood	Network measures	Evolutionary information[*]		
Catalytic site														
Template based	GASS-WEB[64]	http://gass.unifei.edu.br/	Genetic algorithm for matching to templates		✓									Predicts interdomain residues and non-exact template matches.
Machine learning	PREvaIL[65]	NA	Random Forest		✓	✓	✓	✓		✓	✓	*Position specific scoring matrix (PSSM) and Shannon entropy based weighted average percentage calculated using PSI-BLAST.*		
	FunSite [29]	https://github.com/UCL/cath-funsite-predictor	Gradient boosted decision tree		✓	✓	✓	✓	✓	✓	✓	*PSSM and weighted average percentage calculated by alignment within a functional family using default PSI-BLASTt; scorecons value and functional determinant score calculated using groupsim.*	✓	This technique can be used to predict catalytic, ligand binding and protein-protein interface. **Takes care of data imbalance by sampling site to non-site in ratio of 1:6**

	bindPredict[66]	https://github.com/Rostlab/bindPredict	Neural network	✓			✓!					Evolutionary coupling (EC) as calculated by EVcoupling, clustering of the residues with high EC scores and homology based inference from known proteins using BLAST.		This technique can be used to predict nucleic acid binding sites too. For each position calculate a fraction of conservative residues using SNAP2 and EVmutation.
Deep learning	FSCNN[67]	https://simtk.org/projects/fscnn	3D convoluted neural network (best model)		✓									*Based on image classification model, on the entire protein.* The presence of carbon, oxygen, sulphur and nitrogen atoms in each voxel is recorded and is used as an input for the model.
Metapredictor	CSmetaPred_poc[63]	http://14.139.227.206/csmetapred/	Metapredictor based CRpred, CATSID, DISCERN, EXIA2, LIGSite and Fpocket		✓									Combines both catalytic and ligand binding site predictors
Ligand binding site (Small molecule, metal, nucleic acid, peptide) binding site														
Metapredictor	COFACTOR[86]	https://zhanglab.ccmb.med.umich.edu/COFACTOR/	Threading and sequence profile alignment		✓									Also predicts EC number (using BioLip) and GO term (UniProt-GOA and STRING)
Ligand binding site (Small molecule, metal, nucleic acid, peptide) binding site														
Machine Learning	MIonSite[62]	https://github.com/LiangQiaoGu/MIonSite.git	Support vector machine	✓			✓!	✓!				PSSM and Jensen Shannon divergence calculated using default PSI-BLAST.		Differentiates between the different ions. **Takes care of data imbalance by Enhanced AdaBoost**
Other	mFASD[82]	http://staff.ustc.edu.cn/~liangzhi/mfasd/	Structure based analysis of local neighbourhood		✓					✓				Differentiates between the different ions
Small molecule binding site														
Cavity detection	CavityPlus[87]	http://rephar ma.pku.edu.cn/cavityplus	Structure based cavity detection		✓				✓					First detects cavities in proteins and then uses druggability, pharmacophore features, allosteric binding sites of residues and identification of Cys residues in/near binding sites (Cys residues with altered pKa likely to be in/near binding sites) to improve its performance
	CB-DOCK[88]	http://cao.labshare.cn/cb-dock/	Curvature based cavity detection followed by docking		✓				✓					Docks the ligands on the predicted site
Template based	LIBRA-WA[89]	http://www.computationalbiology.it/software.html	Graph theory based structural comparison		✓									The ligand binding sites are clustered based on the similarity in the ligands.
Deep learning	DeepSite[76]	www.playmolecule.org	3D convoluted deep neural network based on atom type and characters		✓									*Based on image classification model; used on entire protein.* Atom based features are used.
	Kalasanty[78]	http://gitlab.com/cheminfIBB/kalasanty	3D convoluted neural network		✓									*Based on image classification models; used on the entire protein.* Atom based features are used. Model was trained to identify deep cavities which are more druggable.
	DeepSurf[79]	https://github.com/stemylonas/DeepSurf.git	3D convoluted neural network		✓								✓	*Based on image classification models; used only on protein surface.* Atom based features are used and the technique produces a ligandability score.
	DeepDrug3D[77]	https://github.com/pulimen	3D convoluted neural network		✓									*Based on image classification models; used only on pocket grids.* For each grid point the statistical

		g/DeepDrug3D											potential for ligand protein interaction are calculated. Model trained only on heme and nucleotide binding sites.
	DeepConV-DTI[75]	https://github.com/GIST-CSBL/DeepConv-DTI	3D Convoluted neural network	✓		✓							Extracted the local residue patterns from the full protein using convoluted neural network and a latent representation of the drug was created using fully connected layers.
	DeepFRI[80]	https://github.com/flatironinstitute/DeepFRI	Graph convolution network	✓	✓	✓							Features are studied as contact maps and residue level features learnt using LSTM-LM. Features generated from 10 million PFAM sequences domains.
	DeepCSeqSite[74]	https://github.com/yfCuiFaith/DeepCSeqSite	Deep convoluted neural network	✓		✓	✓!	✓!			PSSM and Jensen Shannon divergence calculated using PSI-BLAST		Its hierarchical structure captures long distance dependencies and extract low level features.
Machine learning	GRaSP[70]	https://github.com/charles-abreu/GRaSP	Extremely randomized tree		✓		✓			✓			Uses graph-based representation of residue neighbourhood. Multiple classifiers used to make the prediction; the final prediction is based on majority. **Takes care of data imbalance by segmenting the dataset; each segment contains the same binding site and almost equal number of non-binding site.**
	PrankWeb[71]	http://prankweb.cz	Random forest		✓	✓	✓			✓	Jensen Shannon divergence calculated from MUSCLE based MSA	✓	Based on the ligandability of local neighbourhoods of solvent accessible residues
	I-LBR[72]	https://jun-csbio.github.io/I-LBR	Support vector machine	✓			✓!	✓!			PSSM and Jensen Shannon divergence calculated using PSI-BLAST		Has 2 modes – one general and another ligand specific mode. **Takes care of data imbalance by modified random over sampling method.**
Metapredictor	COACH-D[90]	https://yanglab.nankai.edu.cn/COACH-D/	Metapredictor based on S-SITE, COFACTOR, FINDSITE, TM-SITE and ConCavity		✓								Docks the ligands on the predicted site
Peptide binding site													
Template based	SPOT-PEPTIDE[91]	http://sparks-lab.org/tom/SPOT-peptide	Structure alignment		✓								
Machine learning	SPRINT-Str[92]	http://sparks-lab.org/server/SPRINT-Str	Random Forest		✓	✓	✓	✓			*PSSM and entropy calculation using default PSI-BLAST*	✓	Identifies the largest cluster based on Density-based Spatial Clustering of Applications with Noise
	PepBind[93]	http://yanglab.nankai.edu.cn/PepBind/	Support vector machine (SVMPep) and template-based prediction (S-SITE and TM-SITE)		✓		✓	✓		✓			*PSSM, probability matrix, relative entropy and near neighbour correlation coefficient calculated using default PSI-BLAST. HMM based features, probability matrix and near neighbour correlation coefficient calculated using HHblits*		Includes intrinsic disorder based feature
Nucleic acid binding sites													
Machine learning	DNAPred[61]	http://csbio.njust.edu.cn/bioinf/dnapred/	Ensembled hyperplane distance based support vector machine	✓			✓	✓			*PSSM normalized by a standard logistic regression using default PSI-BLAST*		Uses the difference in amino acid frequency between binding and non-binding sites. **Takes care of data imbalance by under-sampling and Enhanced AdaBoost.**
	DRNApred[83]	http://biomine.cs.vcu.edu/servers/DRNApred/	2 layered architecture	✓		✓	✓	✓			Evolutionary profile generated by HHblits		Reduces the cross-prediction between DNA and RNA binding site

	PDRLGB[94]	NA	Light gradient boosted machine learning		✓	✓	✓	✓				PSSM calculated using default PSI-BLAST		Combines both sequence and structure-based features - most other DNA binding site predictors are either sequence or structure based
	iProDNA-CaspNet[95]	NA	Capsule neural network	✓								PSSM normalized by a standard logistic regression using default PSI-BLAST		
	ProNA2020[84]	https://github.com/Rostlab/ProNA2020.git	PSI-BLAST followed by neural network	✓		✓	✓ !	✓ !				Homologues identified using PSI-BLAST as used in PredictProtein server. These profiles were aligned against proteins with GO annotations for DNA/RNA/protein binding		Predicts if a sequence will bind DNA/RNA/Protein and also the binding site residues. For the prior it used natural language processing using ProtVec and also profile kernel SVM, For the latter, it uses neural network.
Other	JET$^2_{DNA}$[85]	http://www.lcqb.upmc.fr/JET2DNA	Non-machine learning based combination of features		✓				✓			Conservation levels calculated using joint evolutionary tree approach. Uses PSI-BLAST to get an MSA	✓	Identification of alternative binding sites on the same protein

Conclusion

In summary, the expansion of protein structure and sequence data, combined with sophisticated approaches for detecting functionally similar homologues and more powerful ML strategies are enabling the development of more accurate functional site prediction tools. Identifying functional homologues remains very challenging particularly in families where structural plasticity enables diverse substrate binding and chemistries. Given the vastness of sequence data, the most powerful strategies for funfam classification primarily exploit that data. Structural data, where available, can help to reduce noise by highlighting predicted sites clustering in 3D. As regards characterising protein families and detecting functional sites, massive expansions in sequence data from metagenome studies mean that for some highly populated families deep learning strategies may be able to learn the specific features associated with the sites driving differences in function.

Cryo-EM and related techniques are expanding the structural data. Recent innovations in protein structure prediction by DeepMind's AlphaFold2 mean that predicted domain structures may soon have comparable quality to experimental structures expanding the structural data further. This, together with other exciting innovations exploiting image-processing and voxellation of structures to capture specific features of functional sites, mean that we can expect significant advances over the next decade.

A challenge that will remain, though, is the sparsity of experimentally characterised sites to train and benchmark the tools. Furthermore, the lack of standard benchmarks for comparing different methods and of recognised assessments (like CASP[96] and CAFA[10–12]) makes it hard to determine which features or strategies bring the most sensitivity and accuracy. Perhaps we need a Critical Assessment of Protein Sites (CAPS)! Initiatives like the PDBe-KB[60] will

help by integrating predicted data to increase coverage and compiling standard benchmark datasets for validating and comparing functional site predictors, and by designing common ontologies for groups depositing the predicted data.

Acknowledgements

NS is funded by BBSRC [BB/S020144/1]; VPW is funded by Wellcome Trust [221327/Z/20/Z]; CR is funded by BBSRC [BB/T002735/1]. NS would like to acknowledge Sanjana Nair, Tejashree Kanitkar and Kaustubh Amritkar for discussions.

Conflict of interest

The authors declare no conflict of interest.

Reference

1. Mi H, Ebert D, Muruganujan A, Mills C, Albou L-P, Mushayamaha T, Thomas PD: **PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API**. *Nucleic Acids Res* 2021, **49**:D394–D403.
2. Haft DH, Selengut JD, White O: **The TIGRFAMs database of protein families**. *Nucleic Acids Res* 2003, **31**:371–373.
3. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, et al.: **Pfam: The protein families database in 2021**. *Nucleic Acids Res* 2021, **49**:D412–D419.
4. Andreeva A, Kulesha E, Gough J, Murzin AG: **The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures**. *Nucleic Acids Res* 2020, **48**:D376–D382.
5. Sillitoe I, Dawson N, Lewis TE, Das S, Lees JG, Ashford P, Tolulope A, Scholes HM, Senatorov I, Bujan A, et al.: **CATH: expanding the horizons of structure-based functional annotations for genome sequences**. *Nucleic Acids Res* 2019, **47**:D280–D284.
6. Morgat A, Lombardot T, Coudert E, Axelsen K, Neto TB, Gehant S, Bansal P, Bolleman J, Gasteiger E, de Castro E, et al.: **Enzyme annotation in UniProtKB using Rhea**. *Bioinformatics* 2020, **36**:1896–1901.
7. **The Gene Ontology resource: enriching a GOld mine**. *Nucleic Acids Res* 2020, **49**:D325–D334.
8. UniProt Consortium T: **UniProt: the universal protein knowledgebase**. *Nucleic Acids Res* 2018, **46**:2699.
9. Brown DP, Krishnamurthy N, Sjölander K: **Automated Protein Subfamily Identification and Classification**. *PLOS Comput Biol* 2007, **3**:e160.
10. Zhou N, Jiang Y, Bergquist TR, Lee AJ, Kacsoh BZ, Crocker AW, Lewis KA, Georghiou G, Nguyen HN, Hamid MN, et al.: **The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens**. *Genome Biol* 2019, **20**:244.
11. 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.
12. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A, et al.: **A large-scale evaluation of computational protein function prediction**. *Nat Methods* 2013, **10**:221–227.
13. Lichtarge O, Bourne HR, Cohen FE: **An Evolutionary Trace Method Defines Binding Surfaces Common to Protein Families**. *J Mol Biol* 1996, **257**:342–358.

14. del Sol Mesa A, Pazos F, Valencia A: **Automatic Methods for Predicting Functionally Important Residues**. *J Mol Biol* 2003, **326**:1289–1302.
15. Sahraeian SM, Luo KR, Brenner SE: **SIFTER search: a web server for accurate phylogeny-based protein function prediction**. *Nucleic Acids Res* 2015, **43**:W141–W147.
16. Sillitoe I, Bordin N, Dawson N, Waman VP, Ashford P, Scholes HM, Pang CSM, Woodridge L, Rauer C, Sen N, et al.: **CATH: increased structural coverage of functional space**. *Nucleic Acids Res* 2021, **49**:D266–D273.
17. Rivoire O, Reynolds KA, Ranganathan R: **Evolution-Based Functional Decomposition of Proteins**. *PLOS Comput Biol* 2016, **12**:e1004817.
18. Mihaljević L, Urban S: **Decoding the Functional Evolution of an Intramembrane Protease Superfamily by Statistical Coupling Analysis**. *Structure* 2020, **28**:1329-1336.e4.
19. Narayanan C, Gagné D, Reynolds KA, Doucet N: **Conserved amino acid networks modulate discrete functional properties in an enzyme superfamily**. *Sci Rep* 2017, **7**:3207.
20. **Coevolution-based inference of amino acid interactions underlying protein function | eLife**. [date unknown],
21. Neuwald AF, Aravind L, Altschul SF: **Inferring joint sequence-structural determinants of protein functional specificity**. *eLife* 2018, **7**.

.. This method not only clusters the MSA into functional families, but it also determines correlated residue-patterns which define each subcluster. Hierarchically subsplitting the MSA into smaller sets allows for multiple levels of functional sub-classification. They tested their approach on a variety of superfamilies showing good results.

22. Lee DA, Rentzsch R, Orengo C: **GeMMA: functional subfamily classification within superfamilies of predicted protein structural domains**. *Nucleic Acids Res* 2010, **38**:720–737.
23. Sillitoe I, Lewis TE, Cuff A, Das S, Ashford P, Dawson NL, Furnham N, Laskowski RA, Lee D, Lees JG, et al.: **CATH: comprehensive structural and functional annotations for genome sequences**. *Nucleic Acids Res* 2015, **43**:D376–D381.
24. Capra JA, Singh M: **Characterization and prediction of residues determining protein functional specificity**. *Bioinformatics* 2008, **24**:1473–1480.
25. Das S, Lee D, Sillitoe I, Dawson NL, Lees JG, Orengo CA: **Functional classification of CATH superfamilies: a domain-based approach for protein function annotation**. *Bioinformatics* 2015, **31**:3460–3467.
26. 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.
27. Lee D, Das S, Dawson NL, Dobrijevic D, Ward J, Orengo C: **Novel Computational Protocols for Functionally Classifying and Characterising Serine Beta-Lactamases**. *PLoS Comput Biol* 2016, **12**:e1004926.
28. Valdar WSJ: **Scoring residue conservation**. *Proteins* 2002, **48**:227–241.
29. Das S, Scholes HM, Sen N, Orengo C: **CATH functional families predict functional sites in proteins**. *Bioinforma Oxf Engl* 2020, doi:10.1093/bioinformatics/btaa937.

. This is one of the few tools that predicts multiple types of functional sites - catalytic, ligand binding and protein-protein interaction site for a single query structure. FunSite improves the prediction performance by combining sequence, structural and 3D network-based properties with residue conservation data derived from functional families.

30. Copp JN, Akiva E, Babbitt PC, Tokuriki N: **Revealing Unexplored Sequence-Function Space Using Sequence Similarity Networks**. *Biochemistry* 2018, **57**:4651–4662.
31. Viborg AH, Terrapon N, Lombard V, Michel G, Czjzek M, Henrissat B, Brumer H: **A subfamily roadmap of the evolutionarily diverse glycoside hydrolase family 16 (GH16)**. *J Biol Chem* 2019, **294**:15973–15986.

32. Knutson ST, Westwood BM, Leuthaeuser JB, Turner BE, Nguyendac D, Shea G, Kumar K, Hayden JD, Harper AF, Brown SD, et al.: **An approach to functionally relevant clustering of the protein universe: Active site profile-based clustering of protein structures and sequences**. *Protein Sci* 2017, **26**:677–699.
33. Harper AF, Leuthaeuser JB, Babbitt PC, Morris JH, Ferrin TE, Poole LB, Fetrow JS: **An Atlas of Peroxiredoxins Created Using an Active Site Profile-Based Approach to Functionally Relevant Clustering of Proteins**. *PLOS Comput Biol* 2017, **13**:e1005284.
34. Lima EB de, Júnior WM, Melo-Minardi RC de: **Isofunctional Protein Subfamily Detection Using Data Integration and Spectral Clustering**. *PLOS Comput Biol* 2016, **12**:e1005001.
35. Bashton M, Chothia C: **The generation of new protein functions by the combination of domains**. *Struct Lond Engl 1993* 2007, **15**:85–99.
36. Lee H, Kim I, Han SK, Kim D, Kong J, Kim S: **Domain-mediated interactions for protein subfamily identification**. *Sci Rep* 2020, **10**:264.

. This subsplits a protein interaction network based on the domain-mediated interactions of the proteins involved. From the thereby generated domain-resolved networks they can then build their protein funfams. This is the first approach using protein multi-domain information in creating functional families.

37. Han J-H, Batey S, Nickson AA, Teichmann SA, Clarke J: **The folding and evolution of multidomain proteins**. *Nat Rev Mol Cell Biol* 2007, **8**:319–330.
38. Zhang S, Li H, Krieger JM, Bahar I: **Shared Signature Dynamics Tempered by Local Fluctuations Enables Fold Adaptability and Specificity**. *Mol Biol Evol* 2019, **36**:2053–2068.
39. Mitchell AL, Almeida A, Beracochea M, Boland M, Burgin J, Cochrane G, Crusoe MR, Kale V, Potter SC, Richardson LJ, et al.: **MGnify: the microbiome analysis resource in 2020**. *Nucleic Acids Res* 2020, **48**:D570–D578.
40. **DeepFam: deep learning based alignment-free method for protein family modeling and prediction | Bioinformatics | Oxford Academic**. [date unknown],
41. Feldbauer R, Gosch L, Lüftinger L, Hyden P, Flexer A, Rattei T: **DeepNOG: fast and accurate protein orthologous group assignment**. *Bioinformatics* 2020, doi:10.1093/bioinformatics/btaa1051.
42. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, Gibbs T, Feher T, Angerer C, Steinegger M, et al.: **ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing**. *bioRxiv* 2020, doi:10.1101/2020.07.12.199554.
43. Littmann M, Bordin N, Heinzinger M, Orengo C, Rost B: **Clustering FunFams using sequence embeddings improves EC purity**. *bioRxiv* 2021, doi:10.1101/2021.01.21.427551.
44. Chakraborty A, Chakrabarti S: **A survey on prediction of specificity-determining sites in proteins**. *Brief Bioinform* 2015, **16**:71–88.
45. Bradley D, Viéitez C, Rajeeve V, Selkrig J, Cutillas PR, Beltrao P: **Sequence and Structure-Based Analysis of Specificity Determinants in Eukaryotic Protein Kinases**. *Cell Rep* 2021, **34**:108602.
46. Joo S, Cho IJ, Seo H, Son HF, Sagong H-Y, Shin TJ, Choi SY, Lee SY, Kim K-J: **Structural insight into molecular mechanism of poly(ethylene terephthalate) degradation**. *Nat Commun* 2018, **9**:382.
47. Suplatov D, Shalaeva D, Kirilin E, Arzhanik V, Švedas V: **Bioinformatic analysis of protein families for identification of variable amino acid residues responsible for functional diversity**. *J Biomol Struct Dyn* 2014, **32**:75–87.
48. Chagoyen M, García-Martín JA, Pazos F: **Practical analysis of specificity-determining residues in protein families**. *Brief Bioinform* 2016, **17**:255–261.
49. Timonina D, Sharapova Y, Švedas V, Suplatov D: **Bioinformatic analysis of subfamily-specific regions in 3D-structures of homologs to study functional diversity and conformational plasticity in protein superfamilies**. *Comput Struct Biotechnol J* 2021, **19**:1302–1311.

50. Suplatov D, Sharapova Y, Geraseva E, Švedas V: **Zebra2: advanced and easy-to-use web-server for bioinformatic analysis of subfamily-specific and conserved positions in diverse protein superfamilies**. *Nucleic Acids Res* 2020, **48**:W65–W71.
51. da Fonseca NJ, Afonso MQL, de Oliveira LC, Bleicher L: **A new method bridging graph theory and residue co-evolutionary networks for specificity determinant positions detection**. *Bioinforma Oxf Engl* 2019, **35**:1478–1485.

. The method uses a co-evolutionary network model for detection of residue sets that tend to co-vary in a well-sampled MSA. The method detects communities of co-evolved residues which could serve as determinants of functional sub-classification in a protein family. The algorithm is also tuned to detect marginally-conserved SDP sites.

52. Fonseca NJ, Afonso MQL, Carrijo L, Bleicher L: **CONAN: A web application to detect specificity determinants and functional sites by amino acids co-variation network analysis**. *Bioinforma Oxf Engl* 2020, doi:10.1093/bioinformatics/btaa713.
53. Tondnevis F, Dudenhausen EE, Miller AM, McKenna R, Altschul SF, Bloom LB, Neuwald AF: **Deep Analysis of Residue Constraints (DARC): identifying determinants of protein functional specificity**. *Sci Rep* 2020, **10**:1691.

. The method applies statistical approaches to investigate subfamily-specific residues and provides visualization of statistically significant 3D clusters formed by them. In addition, the method measures co-variation between residues (i.e. direct couplings) and determines correlations between (i) subgroup-specific pattern residues and direct couplings, and (ii) between 3D structure and direct couplings/subgroup-specific pattern residues. Thus, the method uniquely provides distinct kinds of features in the context of protein specificity and can work for very large sequence datasets as large as half a million sequences.

54. Malinverni D, Barducci A: **Coevolutionary Analysis of Protein Subfamilies by Sequence Reweighting**. *Entropy Basel Switz* 2020, **21**:1127.
55. Pitarch B, Ranea JAG, Pazos F: **Protein residues determining interaction specificity in paralogous families**. *Bioinformatics* 2020, doi:10.1093/bioinformatics/btaa934.
56. **High-Resolution Identification of Specificity Determining Positions in the LacI Protein Family Using Ensembles of Sub-Sampled Alignments**. [date unknown],
57. Chen Z, Liu X, Li F, Li C, Marquez-Lago T, Leier A, Akutsu T, Webb GI, Xu D, Smith AI, et al.: **Large-scale comparative assessment of computational predictors for lysine post-translational modification sites**. *Brief Bioinform* 2019, **20**:2267–2290.
58. Ding Z, Kihara D: **Computational Methods for Predicting Protein-Protein Interactions Using Various Protein Features**. *Curr Protoc Protein Sci* 2018, **93**:e62.
59. He W, Wei L, Zou Q: **Research progress in protein posttranslational modification site prediction**. *Brief Funct Genomics* 2018, **18**:220–229.
60. PDBe-KB consortium: **PDBe-KB: a community-driven resource for structural and functional annotations**. *Nucleic Acids Res* 2020, **48**:D344–D353.
61. Zhu Y-H, Hu J, Song X-N, Yu D-J: **DNAPred: Accurate Identification of DNA-Binding Sites from Protein Sequence by Ensembled Hyperplane-Distance-Based Support Vector Machines**. *J Chem Inf Model* 2019, **59**:3057–3071.
62. Qiao L, Xie D: **MIonSite: Ligand-specific prediction of metal ion-binding sites via enhanced AdaBoost algorithm with protein sequence information**. *Anal Biochem* 2019, **566**:75–88.
63. Choudhary P, Kumar S, Bachhawat AK, Pandit SB: **CSmetaPred: a consensus method for prediction of catalytic residues**. *BMC Bioinformatics* 2017, **18**:583.
64. Moraes JPA, Pappa GL, Pires DEV, Izidoro SC: **GASS-WEB: a web server for identifying enzyme active sites based on genetic algorithms**. *Nucleic Acids Res* 2017, **45**:W315–W319.

65. Song J, Li F, Takemoto K, Haffari G, Akutsu T, Chou K-C, Webb GI: **PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural, and network features in a machine-learning framework**. *J Theor Biol* 2018, **443**:125–137.
66. Schelling M, Hopf TA, Rost B: **Evolutionary couplings and sequence variation effect predict protein binding sites**. *Proteins Struct Funct Bioinforma* 2018, **86**:1064–1074.
67. Torng W, Altman RB: **High precision protein functional site detection using 3D convolutional neural networks**. *Bioinformatics* 2019, **35**:1503–1512.
68. Shoemaker BA, Zhang D, Thangudu RR, Tyagi M, Fong JH, Marchler-Bauer A, Bryant SH, Madej T, Panchenko AR: **Inferred Biomolecular Interaction Server--a web server to analyze and predict protein interacting partners and binding sites**. *Nucleic Acids Res* 2010, **38**:D518-524.
69. Yang J, Roy A, Zhang Y: **BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions**. *Nucleic Acids Res* 2013, **41**:D1096-1103.
70. Santana CA, Silveira S de 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.

. This technique represents each residue and its neighbors as a graph. The physicochemical and topological properties of atoms of a residue and its non-covalent neighbors are modeled as a graph to record the residue environments.

71. 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–W349.
72. Hu J, Rao L, Fan X, Zhang G: **Identification of ligand-binding residues using protein sequence profile alignment and query-specific support vector machine model**. *Anal Biochem* 2020, **604**:113799.
73. Scheibenreif L, Littmann M, Orengo C, Rost B: **FunFam protein families improve residue level molecular function prediction**. *BMC Bioinformatics* 2019, **20**:400.
74. Cui Y, Dong Q, Hong D, Wang X: **Predicting protein-ligand binding residues with deep convolutional neural networks**. *BMC Bioinformatics* 2019, **20**:93.
75. Lee I, Keum J, Nam H: **DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences**. *PLOS Comput Biol* 2019, **15**:e1007129.

. Convoluted neural networks used on protein sequence and convolutions are performed on subsequences of various lengths to capture local residue patterns. The model has been trained on various drug-target interaction databases and optimized using external validation set leading to improved prediction accuracy. Being a sequence based technique, it can be used for a larger number of cases as compared to structure based prediction tools.

76. Jiménez J, Doerr S, Martínez-Rosell G, Rose AS, De Fabritiis G: **DeepSite: protein-binding site predictor using 3D-convolutional neural networks**. *Bioinforma Oxf Engl* 2017, **33**:3036–3042.

. DeepSite was one of the first techniques to consider the structure of the protein as a 3D image with predictions based on properties of the atom types. The method exploits deep convoluted neural networks. The unusual feature is that it is completely independent of the residue type. The descriptors used in this work are general in nature and do not depend on the specific problem being solved, and hence can be extended to other fields of computational structural biology.

77. Pu L, Govindaraj RG, Lemoine JM, Wu H-C, Brylinski M: **DeepDrug3D: Classification of ligand-binding pockets in proteins with a convolutional neural network**. *PLoS Comput Biol*

2019, **15**:e1006718.
78. Stepniewska-Dziubinska MM, Zielenkiewicz P, Siedlecki P: **Improving detection of protein-ligand binding sites with 3D segmentation**. *Sci Rep* 2020, **10**:5035.
79. Mylonas SK, Axenopoulos A, Daras P: **DeepSurf: a surface-based deep learning approach for the prediction of ligand binding sites on proteins**. *Bioinformatics* 2021, doi:10.1093/bioinformatics/btab009.
80. Gligorijevic V, Renfrew PD, Kosciolek T, Leman JK, Berenberg D, Vatanen T, Chandler C, Taylor BC, Fisk IM, Vlamakis H, et al.: *Structure-Based Protein Function Prediction using Graph Convolutional Networks*. bioRxiv; 2019.
81. Zhao Z, Peng Z, Yang J: **Improving Sequence-Based Prediction of Protein-Peptide Binding Residues by Introducing Intrinsic Disorder and a Consensus Method**. *J Chem Inf Model* 2018, **58**:1459–1468.

. This tool is a peptide binding site meta-predictor that combines both template based predictions and machine learning based predictions. The improvements in the machine learning predictor are achieved by using intrinsic disorder characteristics to identify peptide binding sites, as such sites are frequently associated with disorder. In addition, the usage of both sequence and structure based template dependent predictors improve the quality of prediction.

82. He W, Liang Z, Teng M, Niu L: **mFASD: a structure-based algorithm for discriminating different types of metal-binding sites**. *Bioinforma Oxf Engl* 2015, **31**:1938–1944.
83. Yan J, Kurgan L: **DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues**. *Nucleic Acids Res* 2017, **45**:e84.
84. Qiu J, Bernhofer M, Heinzinger M, Kemper S, Norambuena T, Melo F, Rost B: **ProNA2020 predicts protein–DNA, protein–RNA, and protein–protein binding proteins and residues from sequence**. *J Mol Biol* 2020, **432**:2428–2443.
85. Corsi F, Lavery R, Laine E, Carbone A: **Multiple protein-DNA interfaces unravelled by evolutionary information, physico-chemical and geometrical properties**. *PLoS Comput Biol* 2020, **16**:e1007624.

. This DNA binding site predictor exploits evolutionary conservation, curvature and physico-chemical properties of amino acids, without the use of machine learning. It is effective at identifying the binding sites on unbound structures (most DNA binding sites undergo major changes during the binding event) and proteins having multiple binding sites. This technique also has similar performance in RNA binding site predictions, indicating that both DNA and RNA binding sites have similar properties.

86. Zhang C, Freddolino PL, Zhang Y: **COFACTOR: improved protein function prediction by combining structure, sequence and protein-protein interaction information**. *Nucleic Acids Res* 2017, **45**:W291–W299.
87. Xu Y, Wang S, Hu Q, Gao S, Ma X, Zhang W, Shen Y, Chen F, Lai L, Pei J: **CavityPlus: a web server for protein cavity detection with pharmacophore modelling, allosteric site identification and covalent ligand binding ability prediction**. *Nucleic Acids Res* 2018, **46**:W374–W379.
88. Liu Y, Grimm M, Dai W-T, Hou M-C, Xiao Z-X, Cao Y: **CB-Dock: a web server for cavity detection-guided protein-ligand blind docking**. *Acta Pharmacol Sin* 2020, **41**:138–144.
89. 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**. *Bioinforma Oxf Engl* 2018, **34**:878–880.
90. 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.

91. Litfin T, Yang Y, Zhou Y: **SPOT-Peptide: Template-Based Prediction of Peptide-Binding Proteins and Peptide-Binding Sites**. *J Chem Inf Model* 2019, **59**:924–930.
92. Taherzadeh G, Zhou Y, Liew AW-C, Yang Y: **Structure-based prediction of protein– peptide binding regions using Random Forest**. *Bioinformatics* 2018, **34**:477–484.
93. Zhao Z, Peng Z, Yang J: **Improving Sequence-Based Prediction of Protein-Peptide Binding Residues by Introducing Intrinsic Disorder and a Consensus Method**. *J Chem Inf Model* 2018, **58**:1459–1468.
94. Deng L, Pan J, Xu X, Yang W, Liu C, Liu H: **PDRLGB: precise DNA-binding residue prediction using a light gradient boosting machine**. *BMC Bioinformatics* 2018, **19**:522.
95. Nguyen BP, Nguyen QH, Doan-Ngoc G-N, Nguyen-Vo T-H, Rahardja S: **iProDNA-CapsNet: identifying protein-DNA binding residues using capsule neural networks**. *BMC Bioinformatics* 2019, **20**:634.
96. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J: **Critical assessment of methods of protein structure prediction (CASP)—Round XIII**. *Proteins Struct Funct Bioinforma* 2019, **87**:1011–1020.