Towards a comprehensive view of the pocketome universe – biological implications and algorithmic challenges.

Hanne Zillmer and Dirk Walther

Max Planck Institute of Molecular Plant Physiology

Am Mühlenberg 1

14476 Potsdam-Golm

Germany

+49-331-5678 108

zillmer|walther@mpimp-golm.mpg.de

Abstract

With the availability of reliably predicted 3D-structures for essentially all known proteins, characterizing the entirety of protein - small-molecule interaction sites (binding pockets) has become a possibility. The aim of this study was to identify and analyze all compound-binding sites, i.e. the pocketomes, of eleven different species' from different kingdoms of life to discern evolutionary trends as well as to arrive at a global cross-species view of the pocketome universe. All protein structures available in the AlphaFold database for each species were subjected to computational binding site predictions. The resulting set of potential binding sites was inspected for overlaps with known pockets and annotated with regard to the protein domains in which the pockets are located. 2D-projections of all pockets embedded in a 128-dimensional feature space and characterizing them with regard to selected physicochemical properties, provide informative, global pocketome maps that reveal differentiating features between pockets. Our study revealed a sub-linear scaling law of the number of unique binding sites relative to the number of unique protein structures per species. Thus, larger proteomes harbor less than proportionally more different binding sites than species with smaller proteomes. However, this relationship is mainly driven by large numbers of singletons, i.e. binding sites that are not similar to any other binding site in the same species. We discuss the significance of this finding as well as identify critical and unmet algorithmic challenges.

Introduction

To exert their biochemical function, proteins interact with different entities such as small molecules, referred to as compounds, or – in biological systems – metabolites. During this process, compounds interact with cavities on protein surfaces, so-called binding pockets. Protein binding sites have been a subject of both experimental and bioinformatic research for many years, with the aim to assess similarities between them, to identify potential ligands for binding sites, or all potential target proteins for a given compound. A binding pocket is typically defined as all residues within a certain distance to the ligand's heavy atoms in a protein-ligand complex [1]. As this definition requires detailed 3D-structural information of protein-compound complexes, i.e. the position and conformation of the ligand needs to be known, computational approaches have been developed to predict binding pockets from protein 3D-structures alone, i.e. in the absence of an actually bound ligand. Some of these approaches are geometry-based [2,3], while others are energy-based [4,5] or machine-learning-based [6]. With the recent advances in protein structure prediction and the publication of the AlphaFold database, these approaches allow to predict and analyze potential binding sites for whole proteomes of 48 different and evolutionarily diverse species, such as *Arabidopsis thaliana* or human. Thus, a comprehensive and systematic view of compound binding sites on protein surfaces has now become possible.

In this study, we aimed to systematically characterize, cluster, and compare the complete sets of protein-compound binding sites. As in the case of small compound ligands, binding sites are frequently referred to as binding pockets, we are essentially taking aim at describing whole "pocketomes" of different species. By taking a cross-species view and including species of different evolutionary history (bacteria, fungi, plants, mammals), we wished to understand how the pocketomes evolved along with different metabolic needs and organismic capabilities. In particular, as species differ markedly with regard to their proteomes (number of encoded proteins), we wished to discern the corresponding scaling laws that relate the number of unique pockets, reflecting on the potential to bind a corresponding number of unique compounds, to the number of proteins encoded in the respective species. Such analyses may shed light on the trends of evolution with regard of "investment", meaning whether binding capacities scale

proportionally with proteome size, or emerged at increased or decreased rates, with all outcomes being relevant for an understanding of selection processes in evolution.

Each species has a unique set of binding pockets that is involved in metabolic interactions as well as signaling processes. These binding pockets collectively define the species' pocketome. Thus far, only a few studies on pocketomes have been published. These include studies on the pocketome of certain enzyme classes, such as human kinases [7] and G-protein-coupled receptors [8], the pocketome of known 3D structures [9,10], and the species-specific pocketomes, such as *Mycobacterium tuberculosis* [11] and human [12,13]. Most of these studies focused on applications in the field of biomedical research and the druggability of binding sites or discussed differences between drug-protein vs. metabolite-protein binding events [14,15]. An inter-species perspective has not been presented yet.

In order to use known and predicted binding sites for computational comparisons and clustering, they have to be represented in a machine-readable format. Binding site descriptions have been developed at different levels of detail, ranging from coarse-grained representations where each amino acid residue is represented by a single atom, e.g., the C_α atom, to full-atom representations where all atoms of the binding site are used [16]. Other approaches represent binding pockets not based on residue positions but rather as a point cloud [3]. In addition to 3D coordinates, binding pocket representations often include additional information, such as the type of residues aiming to capture the physicochemical properties. Other methods, such as DeeplyTough [17], employ deep learning techniques to learn a representation for a protein binding pocket.

Applying established computational methods that allow for analyses at different levels of structural detail, our study revealed a sub-linear (approximately square root) scaling law of the number of unique binding sites relative to the number of unique protein structures per species. Thus, larger proteomes harbor less than proportionally more different binding sites than species with smaller proteomes. However, this relationship seems to be mainly driven by large numbers of singletons, i.e. binding sites that are not similar to any other binding site in the same species.

We also identified and discuss challenges encountered during the process of performing a global, multi-species computational pocketome analysis and comparison, and propose future

improvements that are in need of being addressed, such as the speed and specificity of binding site comparison algorithms and accounting for co-factors, structural flexibility, and protein-protein interactions.

Materials and Methods

Dataset

Predicted 3D-protein structures of 11 different species covering all kingdoms of life (excluding *Archaea*) were downloaded from the AlphaFold (AF) database v4 (https://alphafold.ebi.ac.uk/) (Table 1) [18]. All predicted structures of all species were considered for compound binding site predictions. Approximately 34% of the investigated protein structures contain at least one transmembrane segment. To ensure the quality of proteins with an alpha-helical transmembrane region, evaluation files from the TmAlphaFold Transmembrane Protein Structure Database (https://tmalphafold.ttk.hu/) were downloaded [19]. All proteins passing less than six quality checks (as reported by the database) were excluded from further analysis, as they are classified as poor (if five quality checks are passed) and failed (if less than five quality checks are passed) models. In addition, all proteins in the AF database that are marked as fragments in UniProt and all proteins with less than 100 amino acid residues were excluded from further analysis.

The potential search space was further reduced by clustering all protein sequences using easy-linclust implemented in MMseqs2 with the default settings [20]. Only the reported cluster representative was used for the inter-species analysis with DeeplyTough (see below).

Additional information, such as known binding sites and domains, was obtained from UniProt [21] and InterPro [22]. The assigned FoldSeek (FS) cluster for each protein was obtained from https://cluster.foldseek.com/ [23].

All ligands found in known pockets were mapped to their respective compound classes using ChEBI ontologies [24]. The ChEBI ID for each ligand was obtained from PDBe-KB and then used as input for a recursive search on ChEBI ontologies. If a ChEBI ID could not be mapped to any of the compound classes, it was assigned as “Other compound”. An overview of all compound classes considered in this study are listed in Table 2.

Compound binding site/ Pocket detection

Potential binding sites were predicted for each investigated protein using P2Rank [6]. P2Rank is a state-of-the-art, machine-learning-based program that was trained to predict potential binding sites via the solvent-accessible surface (SAS) of a protein. After generating a regularly distributed set of points lying on the SAS, feature descriptors are calculated and a ligandability score is predicted for each of the points. Points with a high ligandability are then clustered, and the respective amino acid residues form the predicted binding sites. These pockets are ranked by the cumulative ligandability scores.

In this study, P2Rank was used with its standard settings, as those do not make use of the Debye-Waller B-factor as a feature. For AF-predicted structures, the predicted local distance difference test (pLDDT) is provided in place of the B-factor, which indicates the confidence of the predicted position for each residue and should not be mistaken as B-factors. The resulting potential binding sites for each protein were checked for overlapping residues, i.e. residues assigned to two different sites. If two pockets overlapped, only the pocket that was predicted with a higher probability was kept. To ensure that the predicted pockets have a reasonable size, only predicted pockets that include between 4 and 31 residues were considered further. In addition, pockets with a mean pLDDT < 70 or a mean predicted aligned error (PAE) > 10 were excluded from further analysis, as they are found either in low-quality or intrinsically disordered regions.

The overlap between computationally predicted and the respectively known (annotated) pockets was determined for each protein. If there was any overlap between the amino acid residues of a known and a predicted pocket, the respective known pocket was marked as “found” and the predicted pocket was mapped to the respective compound class corresponding to the ligand that is known to bind to that pocket.

Pocket comparison and dimensionality reduction

Many different binding site comparison tools have been published in recent years. A subset of them has been compared on the ProSPECCTs datasets in [25] and [26]. In this study, we decided to use two different binding site comparison tools, ProBis [27] and DeeplyTough [17].

Other binding site comparison tools were not suitable for this study for several reasons. Tools, such as SiteMine [26] or IChem [28], depend on internal binding site prediction tools, thus making it difficult to compare their performance to other binding site descriptors for predicted binding sites as the predicted sites potentially differ between different prediction tools. In addition, most of them were implemented to compare one pocket to a set of known pockets and not to perform all-against-all comparison. This results in a high average pairwise runtime, ranging between 8 and 28 s, e.g., for SiteHopper [29] and SMAP [30], making them unsuitable for large-scale comparisons [26]. Lastly, the descriptors for some tools, e.g., KRIPO [31] are based on pharmacophore features that are defined for a protein-ligand complex. This makes them unsuitable for predicted binding sites, where the ligand is not known, which is the case for most pockets.

ProBiS [27] compares protein binding sites by detecting common surface structural patches. The solvent accessible surface atoms within the binding site are converted to overlapping subgraphs where each vertex represents one functional group with physicochemical properties, such as hydrogen bond donor or hydrogen bond acceptor. The subgraphs are then compared by constructing a modular product graph that only contains vertices with matching physicochemical properties. For each product graph, a maximum clique is identified which corresponds to the substructure common to the two compared binding sites. This approach allows ProBiS to detect local structure matches of the two compared binding sites and outputs easily interpretable alignments to the user. In this study, ProBiS was used at the level of within species pocket comparisons. As the identification of maximum cliques is computationally expensive, ProBiS is not suitable for inter-species comparisons (11 different species) of all respectively predicted binding sites.

To enable a fast comparison of different pockets via a simple Euclidean distance metric, *DeeplyTough* [17] was employed for the large-scale, inter-species comparisons. DeeplyTough consists of a steerable CNN that is trained to separate dissimilar pockets by a certain margin during the training process. In the first step, a 24 x 24 x 24 Å grid is centered on the defined binding site and each binding site is converted to a 4D tensor with eight channels each one indicating the presence or absence of an atom or pharmacophoric features. These 4D tensors

serve as input to a steerable CNN with six convolutional layers. To achieve rotational invariance, the authors used data augmentation and introduced an additional stability loss. This loss drives the descriptors of the original binding site and its randomly rotated copy towards being the same. The output of DeeplyTough consists of 128-dimensional embedding vectors that can easily be compared and clustered using Euclidean distances.

To plot the resulting 128-dimensional embedding vectors in 2D, different dimensionality reduction techniques were employed. These methods include principal component analysis (PCA), independent component analysis (ICA), uniform manifold approximation (UMAP) and t-distributed stochastic neighbor embedding (tSNE). While for PCA, ICA, and tSNE the implementation in the scikit-learn python package was used [32], the python package umap-learn was used for UMAP [33]. For UMAP, n_neighbors was set to 100 to focus on global shapes and min_dist was set to 0 to enable clustering. For tSNE, the following parameters were used: perplexity = 15, n_iter = 5000, random_state=1, learning_rate="auto", init="random".

Calculation of pocket descriptors

As DeeplyTough does not provide any explicitly interpretable features, several simple pocket descriptors (i.e. properties) were calculated for each pocket to identify the most important features that DeeplyTough is learning. These descriptors include physicochemical properties, such as aromaticity, hydrophobicity, net charge, solvent accessible surface area (SASA), and the structure prediction quality parameters: predicted aligned error (PAE) and predicted local distance difference test (pLDDT).

The aromaticity and hydrophobicity of each predicted pocket was calculated using the ProteinAnalysis module of biopython [34]. This module calculates the aromaticity and hydrophobicity of an amino acid sequence according to [35] and [36], respectively. The net charge of each pocket was estimated using the amino acid residues lining each pocket and the charge of the respective residue type at pH 7.0. All neutral residue types were assigned with a charge of 0 and all negative residue types (Glu, Asp) with a charge of -1. All positively charged residue types (Arg, Lys) were assigned with a charge of +1. Histidine was assigned a charge of +0.5.

The SASA of each pocket was calculated using the Shrake and Rupley algorithm implemented in biopython [34,37]. The sum of all SASA values for each residue in a pocket was obtained and then divided by the maximal solvent accessible surface area for the respective amino acid residue type as defined by [38] to obtain the relative SASA for each pocket.

To check how confident AF is about the relative and local positions of all residues within one binding site, we calculated the mean PAE and the mean pLDDT for each pocket. The PAE values for each protein of interest were obtained from the AF database, and the pLDDT values were extracted from the corresponding protein structures [18].

Spearman's rank correlation coefficients and Pearson correlation coefficients were calculated for all obtained physicochemical and AF-quality pocket descriptors correlated with each dimension of the embedding vector, performed for all dimensionality reduction methods.

Clustering and analysis

As DeeplyTough's embedding space for all binding sites was too densely populated to obtain a meaningful clustering, only ProBiS alignment scores were used to estimate the number of unique pockets per species, as ProBiS operates at an increased level of structural resolution compared to DeeplyTough. The pairwise binding site score matrix for each species was clustered in two ways. In a simple approach, the similarity matrix for each species was clustered by treating it as an adjacency matrix and applying different thresholds for setting edges. Clusters were then detected by determining the number of components detected per graph. In addition, a graph with weighted edges was created based on the similarity matrix using NetworkX [39]. Communities within this graph were then detected using the Leiden algorithm implemented in CDLIB [40,41].

When performing a cross-species analysis of binding site similarity, to estimate the kingdom or species diversity of binding sites mapping to same region in UMAP projection plots, we computed the entropy, *S*, on binned UMAP regions, with

$$S = -\sum_{i=1}^{N} p_i \; log(p_i), \qquad \text{Eq. (1)}$$

where p_i is the fractional frequency of pockets associated with a given kingdom or species relative to all pockets mapping into the same region (bin) of UMAP space, and N is the number considered kingdoms or species, respectively.

Visualizations

All protein visualizations were created using ChimeraX v. 1.7 [42]. All other plots were created with the Plotly library [43].

Results

Pocket identification with P2Rank

A set of 227,492 predicted proteins structures found in the AF database for eleven species including bacteria, plants, unicellular fungi, an insect, and two mammal species (Table 1) was subjected to binding site predictions using P2Rank. To balance the number of true positive and potential false positive binding sites, i.e. sites that are not actually involved in any ligand binding, a threshold of 0.5 was applied to the probability score as predicted by P2Rank for all pockets. This resulted in a set of 75,127 potential binding sites for the eleven different species used in this study. A summary of the known and predicted binding sites of each species is provided in Table 1. The number of predicted pockets is strongly positively correlated with the number of predicted protein structures (Pearson correlation coefficient, r=0.97. Note that at this stage, no protein or pocket clustering was performed. Thus, this correlation pertains to the raw counts.). The largest number of pockets was predicted for SOYBN, which has the largest proteome at the same time, the least pockets are predicted for ECOLI with only 3,834 predicted protein structures. This correlation holds for all species except ORYSJ which has about the same number of predicted pockets as ARATH although it has about 6,000 (~24%) more predicted protein structures than ARATH. Regarding the known pockets, i.a. annotated pockets, the most pockets in UniProt are assigned to ARATH (n=4,016). This is due to the fact that the UniProt database does not only include pockets that were identified based on experimental evidence but also pockets that are assigned based on sequence similarity. Excluding these pockets which are marked by the

Evidence & Conclusion Ontology IDs ECO:0000250, ECO:0000255 and ECO:0000256, reduces the number of known pockets for ARATH to 219 only. Using experimentally annotated pockets only, the largest number of known pockets were found for HUMAN (1,211 pockets) which reflects on the high number of experimentally solved HUMAN protein structures and an annotation bias (increased focus on human proteins and their interactions). Other species, such as MAIZE, seem to be understudied as compared to HUMAN which results in a low number of known pockets although, e.g., MAIZE has the second highest number of predicted structures in this study.

Table 1. **Summary of proteins and pockets for all eleven species used in this study.**

Species	Abbr.	Kingdom	UniProt acc.	#pred. struct.	#known pockets	# known pockets (exp.)	# pred. pockets
Escherichia coli	ECOLI	*Bacteria*	625	3,834	1,025	353	1,675
Saccharomyces cerrevisiae	YEAST	*Fungi*	2311	5,681	1,165	161	2,323
Candida albicans	CANAL	*Fungi*	559	5,741	398	34	2,326
Arabidopsis thaliana	ARATH	*Plantae*	6548	25,218	4,016	219	8,703
Oryza sativa	ORYSJ	*Plantae*	59680	31,293	1,627	34	8,124
Zea mays	MAIZE	*Plantae*	7305	36,282	619	22	10,462
Glycine max	SOYBN	*Plantae*	8827	48,668	1,115	6	16,463
Drosophila melanogaster	DROME	*Animalia*	803	12,605	952	55	4,358
Caenorhabditis elegans	CAEEL	*Animalia*	1940	18,105	1,002	33	6,249
Mus musculus	MOUSE	*Animalia*	589	20,571	3,728	250	7,696
Homo sapiens	HUMAN	*Animalia*	5640	19,494	3,887	1,211	6,901

"# pred. structures" corresponds to the number of protein structures in the AF database without fragments and small proteins. "# known pockets" corresponds to the number of pockets found in UniProt and the "# pred. pockets" to the number of pockets predicted by P2Rank. Note the pocket sets (known/ predicted) are not mutually exclusive.

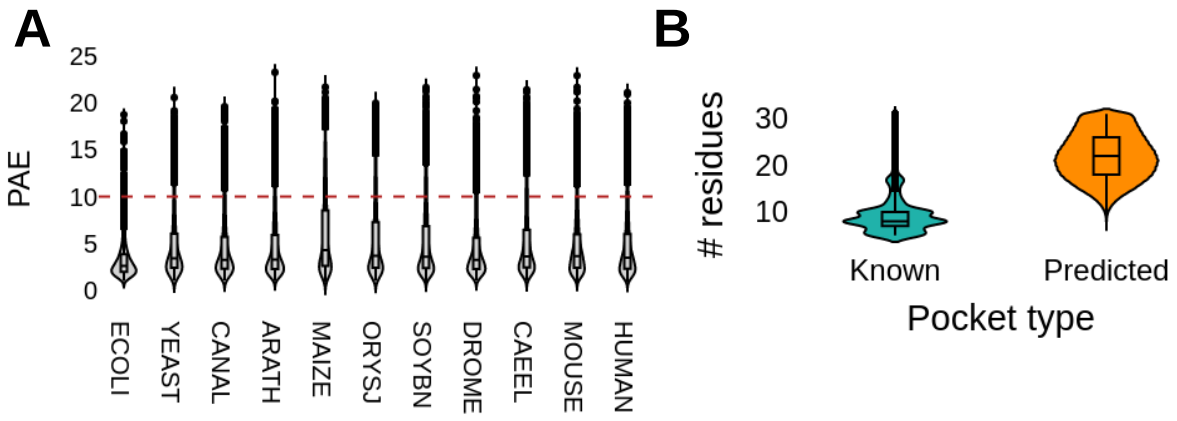


Fig 1. Structural confidence of pocket conformation as judged by AF-Predicted Aligned Error (PAE) and pocket size statistic as gauged by the number of residues lining a given pocket in known and predicted pockets. (A) Violin plots for PAE values per-species. PAE-values correspond to the average of the square pairwise all-against-all residue PAE matrix. All pockets above the red dotted line were not considered for further analysis. The median for all PAE distributions varies between 2.6 and 4.3. MAIZE shows the largest variation in PAE values. (B) On average, predicted pockets consist of more residues than experimentally annotated (known) pockets. While the median number of residues for known pockets is 8, the median number of residues is 22 for predicted pockets.

First, we wished to verify that the quality of the protein structures is comparable across the species used in this study. The distribution of PAE (AF-predicted aligned error) values associated with all pockets of the 11 species and the applied threshold value of 10 Å is visualized in Figure 1A. The largest variation in PAE values was observed for MAIZE, ORYSJ, and SOYBN. These species have the largest proteomes and the most predicted pockets. The least variation was observed for ECOLI, suggesting a lower fraction of less defined, disordered regions in ECOLI. The median PAE ranges between 2.6 and 4.3 Å for all species. Despite the noted range differences, there are no significant differences of the estimated prediction quality between the eleven species, rendering them all suitable for a systematic cross-species comparison.

If available for a given protein, the domain information was mapped onto the respective, predicted binding pockets. While ~23,400 (31 %) pockets could not be assigned to any domain, most of the pockets that have a domain assigned to them are annotated as protein kinase domains (InterPro: IPR000719, 4,234 pockets) and GPCR, rhodopsin-like, 7TM domain (InterPro: IPR017452, 1,574 pockets). A total of 33,367 pockets predicted by P2Rank for all species are inter-domain binding sites that are assigned to two or more different domain types. Most of these pockets are at the interface between two different domains, but 10,339 pockets are found at the

interface of more than two domains. Another subset includes pockets that were found at a "domain-no-domain" interface (27,634 pockets). A small subset of 604 pockets was found in domains of unknown function (DUF).

Figure 1B shows the number of residues per binding site, i.e. residues lining the respective pocket reflecting on binding site size, for known and predicted pockets. Predicted pockets are on average significantly larger (median number of residues=22) than known binding sites (median=8). While known pockets are derived from experimentally solved structures, P2Rank predicts binding sites by clustering potentially druggable surface atoms. Therefore, residues that are not binding to any ligand are potentially included in the predicted pocket, as well. In addition, P2Rank does not include any additional pocket segmentation, resulting in large pockets if several ligands are binding to a similar region, as is, for example, the case, for ATP-dependent reactions.

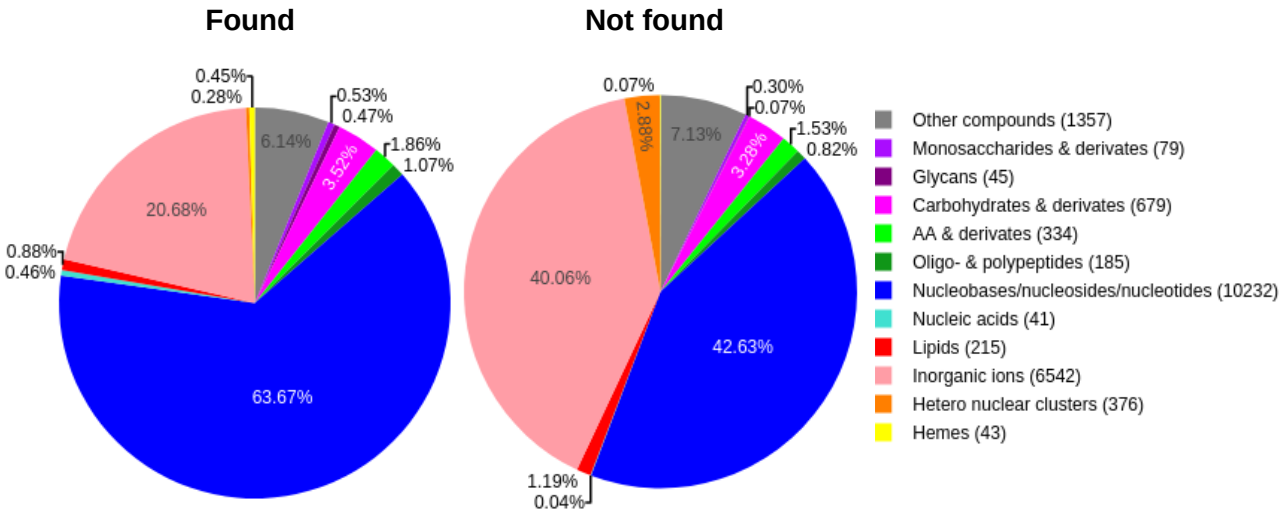


Fig 2. Ligand annotation of pockets.

Known (i.e. annotated) pockets identified by P2Rank used as the binding site predictor and with correspondingly available ligand annotation were classified as "found" (n=7,688, 39.4 %), and "not found" (n=11,820, 60.6 %), otherwise. The total number of occurrences for each compound class is denoted in brackets in the legend.

Exploring the reliability of P2Rank pocket predictions by checking the overlap of predicted sites with known sites, designating them as "found" if at least one residue-based overlap was detected, and "not found" otherwise, we observed that the most "found" as well as "not found" pockets were annotated as binding "nucleobases/ nucleosides/ nucleotides". A 2.4-fold enrichment (p-value < 0.05, Fisher's exact test) for known pockets that are denoted as "found" was observed for this ligand class as compared to all other ligand classes. The high fraction of not identified "nucleobases/ nucleosides/ nucleotides" pockets includes inter-chain pockets as

wells as pockets for which the predicted pendant exceeded the threshold applied to the number of residues. The second largest fraction for both categories is pockets that are annotated as inorganic ions (Fig 2). For inorganic ions, we observed a significant depletion with p-value < 0.05 and odds ratio = 0.39 (Fisher's exact test) for known pockets that are denoted as "found" as compared to all other ligand classes. The binding sites of inorganic ions are often formed by a few amino acid residues, via charge-driven complex formation or covalently, thereby exhibiting less of the hallmark features of compound binding sites (surface invaginations). This makes it difficult to predict their location if they are bound to a site that is distant from any other binding site. Thus, inorganic ions are proportionally overrepresented in the "not found" set. For other ligand classes, there are fewer known binding sites annotated in UniProt. These classes include heme, lipids, carbohydrates and derivatives, and amino acids and derivatives. No significant enrichment or depletion was observed for "Carbohydrates and derivatives" as well as "Amino acids and derivatives", "Oligo- and polypeptides", "Lipids", and "Hetero nuclear clusters" using Fisher's exact test. A summary of the respective statistics of the Fisher's exact test and the number of pockets that were denoted as "found" and "not found" for all ligand classes can be found in Table 2.

Table 2. Summary table for the compound classes of the ligands binding to known pockets.

Compound class	ChEBI ID	# Pockets ("Found")	# Pockets ("Not found")	Odds ratio	p_{adj}-value
Nucleobases, nucleosides & nucleotides	CHEBI:18282, CHEBI:33838, CHEBI:36976	4,999	5,233	2.36	0.0
Inorganic ions	CHEBI:36914	1,624	4,918	0.39	0.0
Other compounds	-	482	875	0.85	0.033
Amino acids & derivatives	CHEBI:33709, CHEBI:35238, CHEBI:83821	334	146	1.22	0.276
Carbohydrates & derivatives	CHEBI:16646, CHEBI:63299	276	403	1.07	1.0
Oligo- & polypeptides	CHEBI:25676, CHEBI:15841	84	101	1.29	0.276
Lipids	CHEBI:18059	69	146	0.74	0.171
Monosaccharides & derivatives	CHEBI:35381, CHEBI:63367	42	37	1.78	0.051
Glycans	CHEBI:37163, CHEBI:167559	37	8	7.26	0.0
Nucleic acids	CHEBI:16991, CHEBI:33697	36	5	11.3	0.0
Hemes	CHEBI:30413	35	8	6.87	0.0
Hetero nuclear clusters	CHEBI:33733	22	354	0.09	0.095
No known compounds	-	-	-	-	-

For all compound classes considered in this study, the respective ChEBI IDs, the number of pockets identified ("found") and not identified ("not found") by P2Rank and the corresponding statistic of the Fisher's exact test are shown. For Fisher's exact tests, each compound class was tested for enrichment in found pockets compared to all other compound classes. The odds ratios signify the fold increase or decrease of detecting sites for the respective class from a 2x2 contingency table of observations (found/not-found x is-class/other-class). The p-values for each compound class were corrected for multiple testing using the Benjamini-Yekutieli procedure to control the false discovery rate.

Several factors may explain why known pockets might not be predicted by P2Rank. One reason is that flexible loops flanking the binding site do not have a stable conformation. These flexible loops are often missing in experimental structures. In AF, they are assigned to a single conformation with a mostly low pLDDT values. An example for this case is shown in Fig 3A. While the two flexible regions are missing completely in the experimental structure of tubulin

polyglutamylase TTLL7 (PDB ID: 4YLR), the ADP binding site is not accessible in the AF predicted structure (UniProt ID: Q6ZT98). Other known binding sites, such as the LacNAc binding site on galectin-3, only consist of few residues that form a flat binding surface (see Fig 3B).

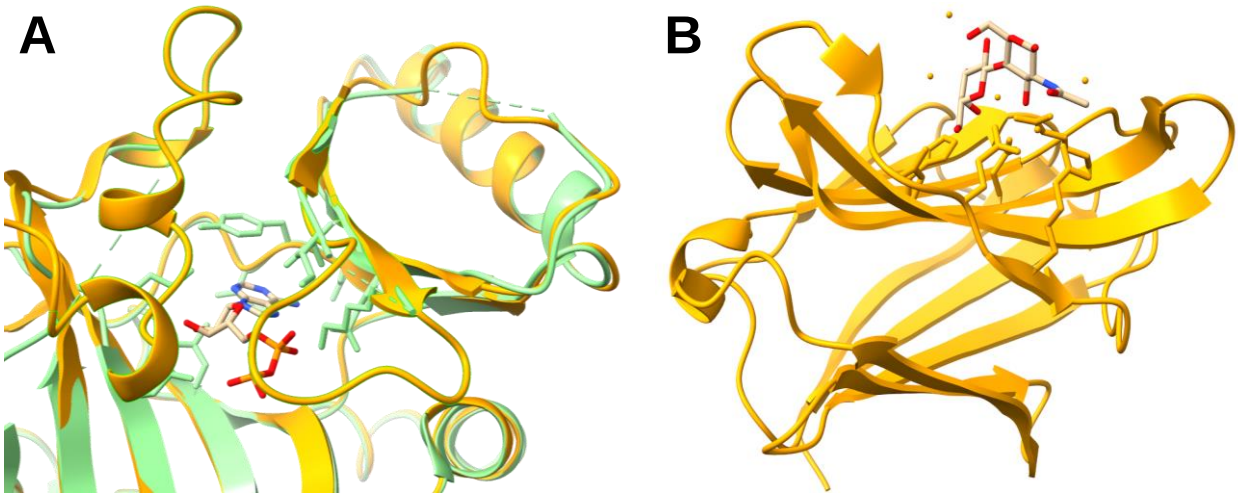


Fig 3. Examples of binding sites that could not be identified by P2RANK due to flexible regions and flat surfaces.

(A) Experimental (PDB ID: 4YLR, orange) and AF structure for Q6ZT98. In 4YLR, ADP (beige) is bound as a ligand. This ADP binding site is blocked in the AF structure (green) by two flexible regions (residues: 161-172, 235-269) that are not present in the experimental structure. Therefore, P2Rank cannot predict the ADP binding sites. (B) Galectin-3 (PDB ID: 1KJL, orange) binds LacNAc (beige). The binding site is found on a flat surface and consists of only few AAs.

Per-species binding site comparison (ProBiS)

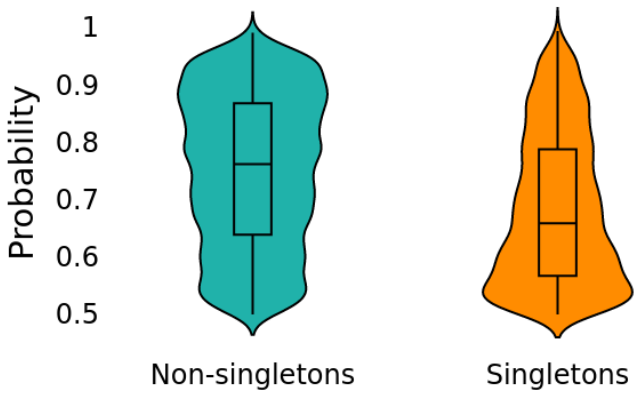


Fig 4. Annotation confidence of pockets for non-singletons and singletons.

Violin plots of the P2Rank pocket probabilities for singletons and non-singletons. The median of the probabilities for non-singletons is 0.76 which is 0.11 higher than the median for the singletons.

As we wished to discern the underlying scaling laws linking the pocketomes of different species to their respective proteomes as well as evolutionary context, we performed per-species comparison of the predicted binding sites using ProBiS to determine the number of unique binding

sites present in each species. This graph-based comparison algorithm results in a sparse similarity matrix, where zeros denote pocket pairs that could not be aligned and thus are considered dissimilar. The higher the alignment scores, the more similar two pockets are. Pockets that could not be aligned to any other pocket of the respective species are termed “singletons”. The number of singletons and the number of communities detected by the Leiden algorithm are listed in Table 3. The fraction of “number of singletons” over the “number of predicted” pockets is smaller for species with larger proteomes and more predicted pockets than for species with smaller proteomes, such as ECOLI, YEAST, and CANAL. For all three species, the fraction of singletons over the number of pockets is greater than 0.5. The lowest fraction of singletons over number of pockets was detected for SOYBN with 0.139. To answer the question whether singletons might be falsely predicted binding sites that are not actually involved in any ligand binding, we compared different properties of the pockets for singletons and non-singletons. Although all of the properties show significant differences between singletons and non-singletons with p-value < 0.01 (t-test), the effect sizes calculated using Cohen’s d differ markedly between different properties. The largest effect size was observed for the probability of the pockets, i.e. the confidence of the pocket prediction. The median probability of the predicted pockets drops from 0.76 for the non-singletons to 0.65 for singletons (Fig 4). In contrast to this medium effect of 0.559, there are only small effects for the other properties with effect sizes varying from 0.024 to 0.337 (see S1 Fig). This implies that singletons of low probability potentially are pockets that are not actually involved in any ligand binding.

Next, we addressed the question as to whether expanding proteomes during the course of evolution commensurate with proportionally increased pocketome sizes. While a proportional increase may be the naive expectation, alternatively, relatively more or fewer pockets emerge during evolution than is expected purely based on proportionality. That is, we wished to understand whether evolution “invests” into new proteins to expand the set of divers compounds that can be processes (via binding to proteins), or whether functions other than compound binding are invested in. If the latter, with more unique protein folds, fewer than proportionally more binding pockets would be found with increased proteome sizes. By contrast, evolution could also proceed “economically” by reusing the same fold but evolving new and additional binding sites on them

via selected point mutations. Under that scenario, more than proportionally more pockets per protein fold would be seen. To explore this relation between unique binding sites and unique folds per species, we plotted the number of unique binding sites (communities + singletons) against the number of FoldSeek (FS) clusters (https://cluster.foldseek.com/) [23] and the number of clusters (only communities) against the number of FS clusters. We decided to gauge the proteome size as the number of FS clusters and not by the number of encoded proteomes to avoid wrongly associating different sizes of protein families with protein "novelty". In addition, this approach accounts for the effect of gene duplication in some plants, such as SOYBN, where 73.6 % of the genes are duplicated according to its BUSCO score. Note that unlike reported above, where we stated a high linear correlation of numbers of proteins and pockets, here we performed a clustering on both entities to really capture unique proteins and pockets.

As shown in Figure 5A, there is a negative correlation between the number of FoldSeek clusters, N_{FS}, in the different species and the number of pockets clusters per FS cluster, $N_{Pocket\ clusters}/N_{FS}$, where $N_{Pocket\ clusters}$ is the number of pocket clusters in a given species. Compared to ORYSJ, which has the lowest ratio of pocket clusters to FS clusters, the ratio between the number of pocket clusters and FS clusters is three times higher for ECOLI, which has only about 1/100th the number of FS clusters compared to ORYSJ. Note that under pure proportionality, we would expect this relationship to be a constant (as we normalize the number of pocket clusters by the number of FS clusters). Regressing $\log(N_{\text{Pocket clusters}})$ as a linear function of $\log(N_{FS})$, a strong linear relationship (Pearson correlation coefficient r= 0.94) was obtained, suggesting a power law relationship with proportionality of $N_p \sim N_{FS}^a$, and a= 0.548 (slope of the regression line). A smaller-than-one exponent a suggests a less than linear (sub-linear) increase of the number of pocket clusters with increasing number of FS clusters, i.e. larger proteomes do not harbor a proportionally greater diversity of binding sites. In fact, the obtained scaling law suggests a proportionality close to the square root (a=0.5) of the number of FS clusters. Of note, when placing N_o objects randomly into a high dimensional features space, a proportionality of the number of resulting clusters, N_c, relative to the number of objects, N_o, would also be expected to yield a power low relationship with a<1, as the density of objects increases with N and sub-linearly more clusters are detected with an increasing number of objects (assuming constant clustering criteria). Thus, upon first

glance, the evolutionary “exploration” of structure space is consistent with a random expansion, but not with a “targeted” expansion, where “targeted” should be understood as the result of selection, as the latter would be signified by larger coefficients, *a*.

However, this sub-linear (near square-root) scaling law between the number of pocket clusters and the number of FS clusters appears to be strongly influenced by the number of singletons, as there are more singletons than communities for all species. Excluding singletons yields a proportional increase (zero slope in Fig 5B) of pocket clusters (now detected and referred to as communities (see Materials and Methods) per FS cluster vs. FS clusters. Note again that in Fig 5B the number of pocket communities on the y-axis was normalized by the number of FS clusters. Excluding singletons, we determined a power law exponent of a= 0.97, i.e. very close to one, suggesting direct proportionality between $N_{Communities}$ *and* N_{FS}. Despite the overall proportionality, there is substantial scatter in the data, with the maximum y-value ($N_{Communities}/N_{FS}$) observed for SOYBN being twofold greater than the value observed minimum of the 11 species (ORYSJ). The differences between MAIZE, ORYSJ, and SOYBN, which all have approximately the same number of FS clusters, can be explained by the low number of singletons per pockets for SOYBN and a lower ratio of predicted pockets per protein for MAIZE and ORYSJ with 0.288 and 0.26, respectively (S1 Table). The latter seems to be, at least partially, driven by a higher fraction of low-quality or disordered protein structures without any predicted binding sites that have a mean pLDDT ≤ 70 as shown in S2 Fig.

Thus, assuming all singleton binding sites to be real, larger proteomes with more FoldSeek clusters evolve sub-linearly more distinct binding sites. However, when inspecting binding site clusters with two or more members (communities), we arrive at the conclusion that larger proteomes do indeed evolve proportionally more and different unique binding sites and, presumably, associated increased processing capacity. And because random processes would result in sub-linear correlations, a linear correlation can be seen as evidence of selection favoring the addition of “novel information”.

Table 3. Pocket clustering statistics.

Species	# Communities	# Singletons	# Singletons/ # Pockets	Largest community
ECOLI	181	932	0.562	75
YEAST	215	1,252	0.547	81
CANAL	206	1,285	0.567	70
ARATH	808	1,952	0.23	748
MAIZE	974	2,204	0.22	940
ORYSJ	701	2,019	0.255	801
SOYBN	1,387	2,207	0.139	1,556
DROME	364	2,014	0.474	149
CAEEL	476	2,304	0.377	284
MOUSE	659	2,296	0.303	684
HUMAN	653	2,373	0.359	336

Communities, i.e. sets of similar pockets, were found by applying the Leiden algorithm. Singletons are lone pockets, not similar to any other pocket in a given species. While the number of singletons is getting larger, the number of singletons per pocket is getting lower the larger the proteome of a species is. The highest fraction is observed for CANAL, ECOLI, and YEAST. These species have the lowest number of communities. The lowest fraction is observed for MAIZE and SOYBN. “Largest community” indicates the number of member pockets in the largest found community.

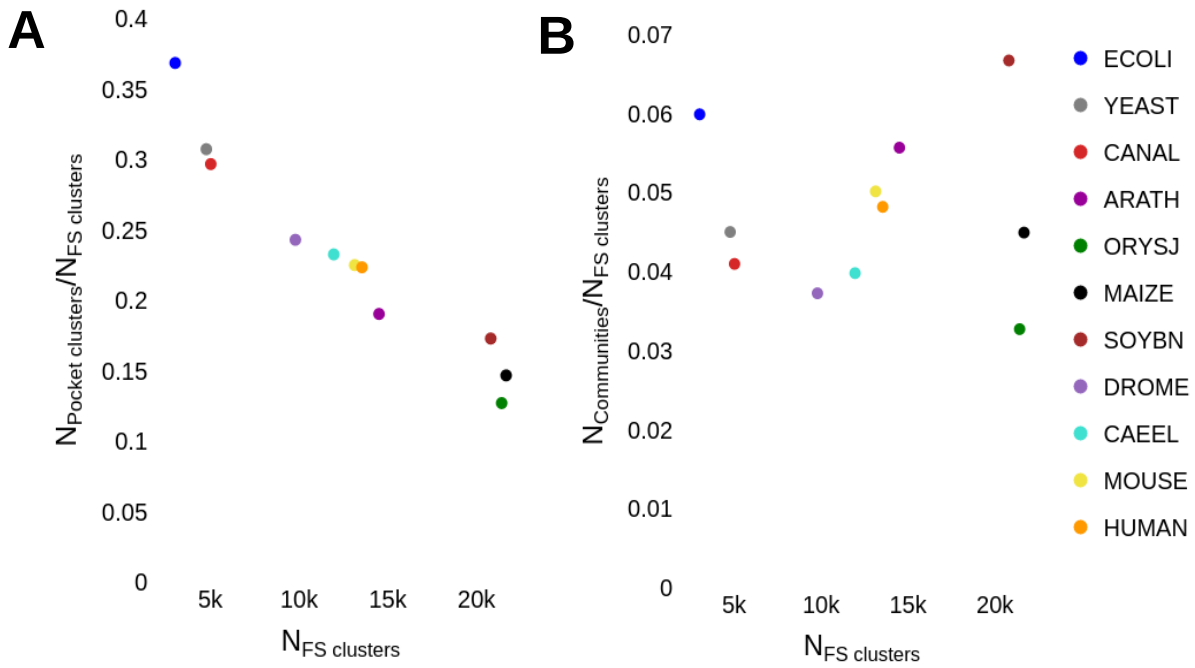


Fig 5. Scaling law of the number of unique pockets vs. proteome size.

(A) Number of unique pockets per species ($N_{Pocket\ clusters}$) normalized by the number of FoldSeek (FS) clusters vs. the number of FS clusters ($N_{FS\ clusters}$). On a log-log scale for $\log(N_{Pocket\ clusters}) \sim \log(N_{FS\ clusters})$, a linear regression yielded: slope = 0.548, intercept = 2.682, Pearson r = 0.959. (B) Number of pocket communities per species, $N_{Communities}$, i.e. without considering singleton pockets, normalized by the number of FoldSeek (FS) clusters vs. the number of FS clusters ($N_{FS\ clusters}$). On a log-log scale for $\log(N_{Communities}) \sim \log(N_{FS\ clusters})$, a linear regression yielded: slope = 0.967, intercept = -2.767, Pearson r = 0.951). Note that the reported slope values correspond to the power law exponent, *a*, discussed in the main text.

Regarding the sizes of the communities, the Leiden algorithm detects a few large communities for each species and many smaller communities that only consist of two or three different pockets (Fig 6). Overall, linear relationships are observed in the log-log plot of cluster-frequency and cluster size, suggesting an overall power-law relationship for the cluster-size distribution, which is oftentimes observed for such data types in biological systems [44]. The largest communities are found in SOYBN, MAIZE, and MOUSE. As ATP/ADP is the most common known ligand and protein kinase is the most common domain for the predicted pockets, it is expected that the largest community is enriched for protein kinase domain annotations. This is the case for all species, except ECOLI, DROME, and MOUSE. For all other species, 85.2 to 97.2 % of the pockets in the largest cluster are found at least partially in a protein kinase domain. The paucity of protein kinase annotations in ECOLI is expected as it does not have the eukaryotic protein kinase domain. The largest proportion of pockets are found in the “no domain”-annotation (15 pockets) for ECOLI. For DROME and MOUSE, the protein kinase domain dominates in the second largest cluster with 144 and 360 members, respectively. The largest community for MOUSE contains mainly pockets that are found in the GPCR, rhodopsin-like, 7TM domain. The largest community for DROME contains mainly pockets that are found in ATPase domains.

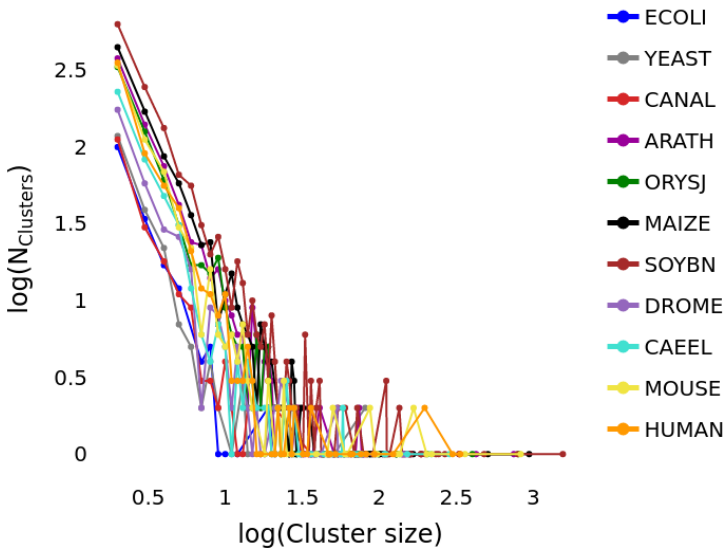


Fig 6. Cluster size frequency distribution for 11 species used in this study.

The log(Cluster size) plotted against the log($N_{Clusters}$). The clustering for all species reveals that there are many clusters that only contain few pockets and only few clusters with many pockets. The largest clusters are found in SOYBN, MAIZE, MOUSE.

Large-scale, inter-species binding site comparison (DeeplyTough)

Thus far, we clustered pockets and proteins per species. Evidently, the pocket datasets could also be inspected across all species. As the computational resources needed for an all-against-all comparison using ProBiS are prohibitively high, we decided to use the embedding vectors obtained by DeeplyTough for a large-scale, inter-species comparison. By using DeeplyTough, a method to describe pockets as numeric vectors of length 128, we were able to perform an all-against-all pocket comparison of all 67,406 pockets across all eleven species used in this study, arriving at a global view of the pocketome “universe”. We present this view primarily by using UMAP as an appropriate projection method, but included tSNE, PCA (principal component analysis), and ICA (independent component analysis) as well (see S3 - S9 Fig).

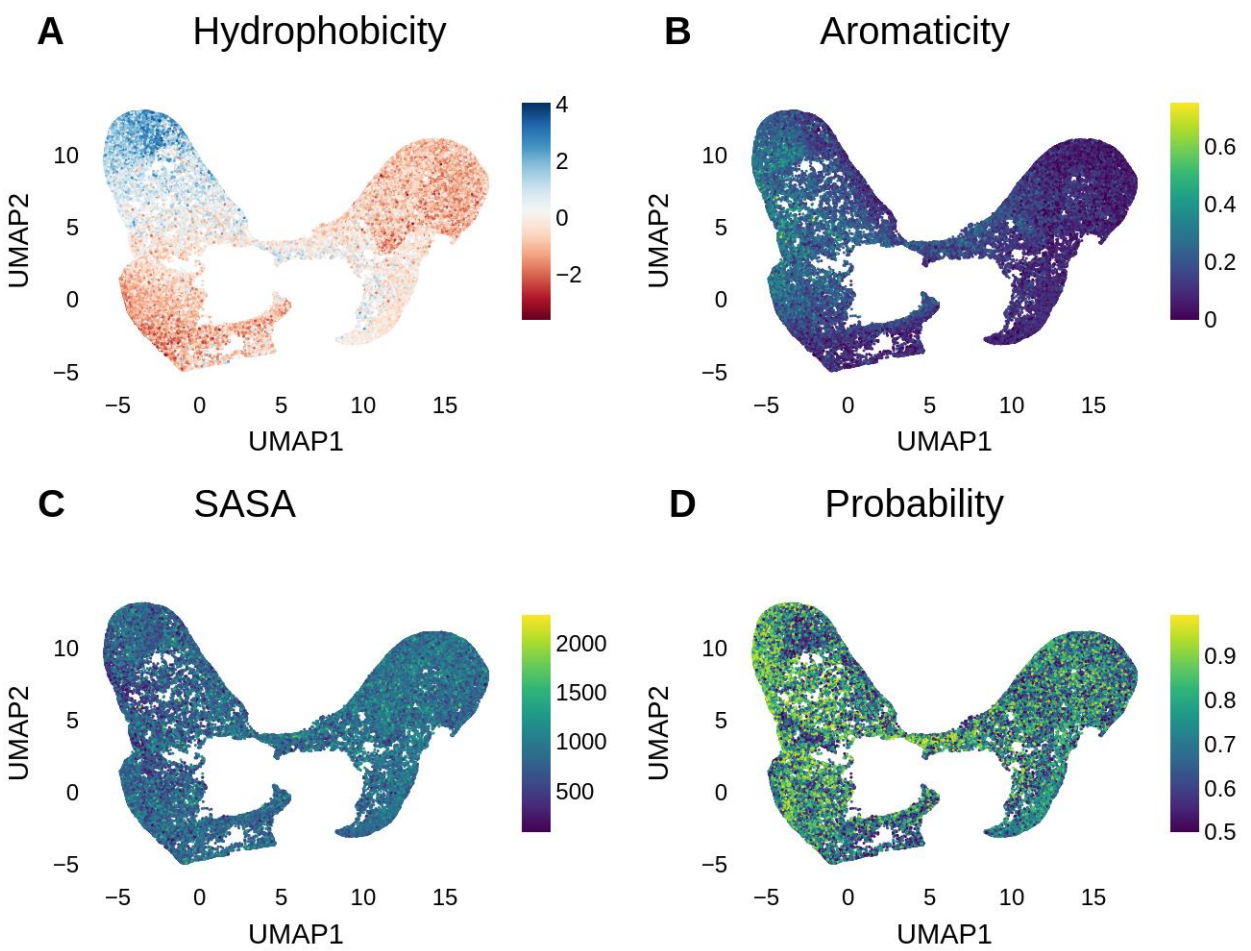


Fig 7. UMAP plots for all predicted pockets of all 11 species colored by different properties.

DeeplyTough seems to be able to differentiate pockets based on their hydrophobicity and aromaticity. While other properties such as SASA and probability seem to be less important. The latter is expected as both properties are no explicit input feature of DeeplyTough.

Table 4. Spearman correlation between different pocket features and the first two dimensions of different dimensionality reduction techniques.

Pocket feature	UMAP1	UMAP2	tSNE1	tSNE2	PC1	PC2	IC1	IC2
Aromaticity	-0.613	-0.005	0.176	-0.086	0.272	0.584	-0.342	0.04
Hydrophobicity	-0.375	0.314	0.217	-0.083	0.709	-0.084	0.379	-0.409
Net charge	-0.124	0.07	-0.004	-0.062	-0.064	-0.018	0.006	0.079
SASA	0.137	0.072	0.01	0.033	0.031	-0.198	0.114	-0.307
Probability	-0.108	0.004	-0.038	-0.028	0.055	0.164	-0.031	0.195
pLDDT	-0.168	-0.162	0.028	0.038	0.123	0.25	-0.246	0.168
PAE	-0.155	0.119	-0.027	-0.041	0.039	-0.168	0.15	-0.082
number of residues	0.184	-0.052	-0.13	0.02	-0.113	-0.087	0.041	0.093

As judged based on absolute Spearman correlation levels, aromaticity, and hydrophobicity are the most important features to discriminate pockets in the tested embedding space dimensions of DeeplyTough for all four dimensionality reduction techniques. The other pocket features are less important.

Figure 7 shows all pockets projected into the first two UMAP components colored by the different pocket properties: hydrophobicity, aromaticity, relative SASA, and probability. Each point in the plot represents one unique pocket representative (representative with regard to proteins of similar sequences as determined by MMseq2, see Materials and Methods). The coloring by properties allows for gauging the most important properties by which the pockets differ. To arrive at a more quantitative assessment, the Spearman correlation for the different dimensionality reduction techniques and their respective main components and features were computed and are listed in Table 4. The UMAP plot and the correlation values show that hydrophobicity (Fig 7A) and aromaticity (Fig 7B) are the most important features for DeeplyTough to discriminate between pockets. This observation holds true for all different dimensionality reduction techniques tested. For the other features, lower correlation values were observed. This is to be expected for “probability” (Fig 7D), pLDDT, and PAE (see S3 Fig) as they are not an explicit feature for neither P2Rank nor DeeplyTough.

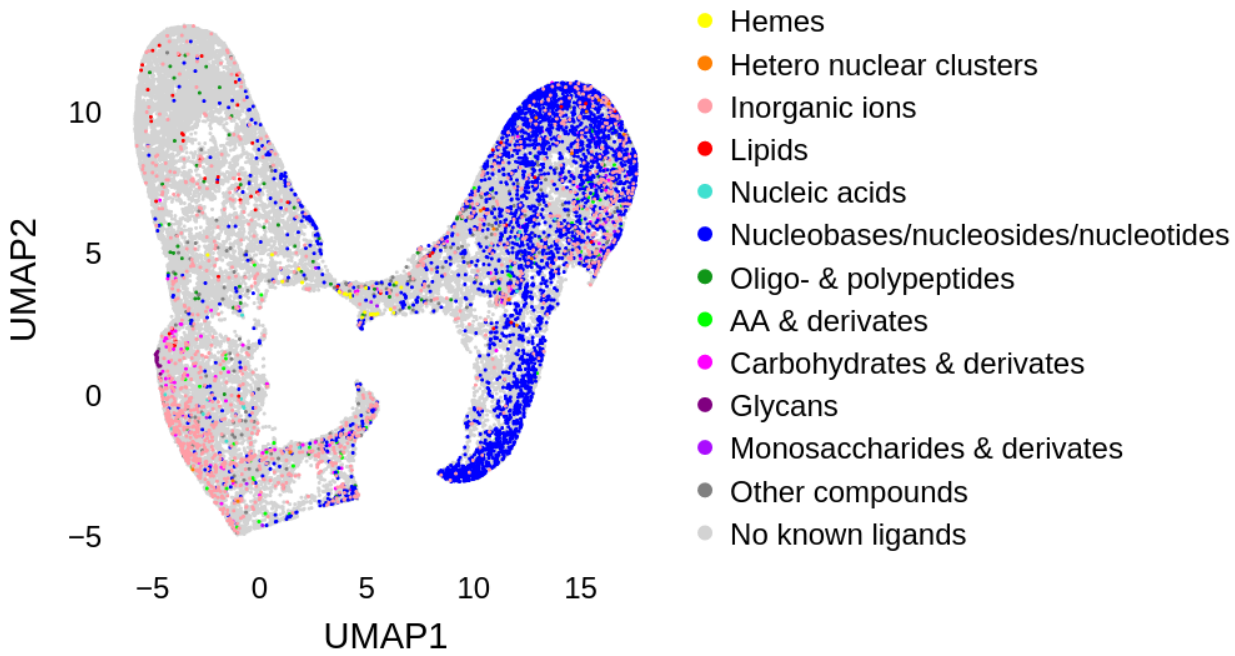


Fig 8. UMAP plot of all pockets of all 11 species colored by different ligand classes.
The most abundant ligand class are nucleobases/nucleosides/nucelotides which are mostly found on the left side of the UMAP plot. While glycan binding sites cluster well, the binding sites for other ligand classes are found in different regions of the UMAP plot.

To examine whether particular compound classes correspond to binding sites displaying characteristic properties, we examined the pocketome projection with regard to their annotation, i.e. which compounds are known to bind to it, provided this information is available (Fig 8). The most abundant ligands found in known pockets are “nucleobases, nucleosides, and nucleotides”, such as ADP and ATP. The predicted pockets that overlap with these pockets are mostly found on the right side of the UMAP plot. As evident from Figure 7, these pockets are hydrophilic to neutral on the hydrophobicity scale and are predicted with moderate to high probability. In addition, they contain few aromatic amino acid residues and are larger than the average predicted binding site (Fig 7 and 12). As mentioned above, P2Rank does not perform any sub-sampling of the predicted pockets. Binding sites that contain several ligands are therefore predicted as one large pocket instead of several separate pockets.

Glycan binding sites are rather hydrophilic, contain moderate amounts of aromatic amino acid residues, and are all found in the Glycoside hydrolase family 16 domain (InterPro: IPR000757). They are found on the lower left side (at the outer border, “west coast”) of the UMAP

plot and cluster well. One reason for that might be their unique shape as they are formed by a set of anti-parallel β-sheets and their connecting loops (example shown in Fig 9). In the vicinity of these pockets, there are several other, not annotated pockets that are found on the glycoside hydrolase family 16 domain. This suggests that more glycan binding sites could be potentially identified with this approach.

For other ligands, such as lipids, carbohydrates, and amino acid derivatives, the predicted binding sites that overlap with known binding sites of those compounds are not clustering strongly. Their associated binding sites are found in several parts of the UMAP plot.

In the upper left part of the UMAP map there is a region of hydrophobic and aromatic pockets that are only overlapping with few known pockets. As compared to other predicted pockets, they are smaller and contain only around 10 to 15 residues (Fig 1B). This region, spanned by UMAP1 values ≤ 1.5 and UMAP2 values ≥ 8, contains about 10,000 pockets and corresponds to the region of hydrophobic pockets in Fig 7A. About 4,000 pockets in this region have a hydrophobicity of greater than 2 according to the Kyte-Doolittle scale. Regarding these highly hydrophobic pockets, about 70% of them are found in proteins that contain a transmembrane domain. While the largest fraction, 1,300 pockets of this subset could not be assigned to any domain, the rest is mostly found on various transporter domains, such as IPR005821, IPR006153 and IPR011527. The other 30% of hydrophobic pockets that are found

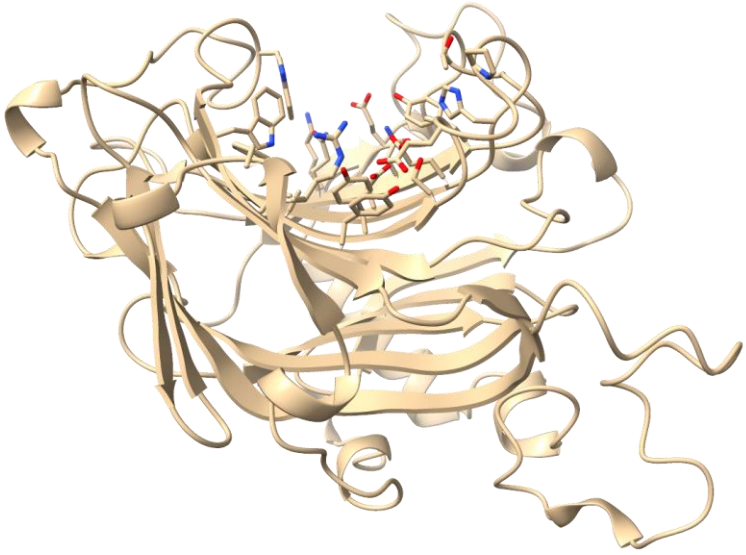


Fig 9. Glycan binding site of Q8LC45.

The binding site residues are displayed as sticks, the rest of the protein as cartoon. The binding site is formed by several anti-parallel β-sheets. This protein visualization was created using ChimeraX (v. 1.7).

in proteins that do not contain any transmembrane domains, could either be not assigned to any domain (405 pockets) or are found in a variety of different domains with the EF-hand domain (InterPro: IPR002048) being the most abundant domain (52 pockets).

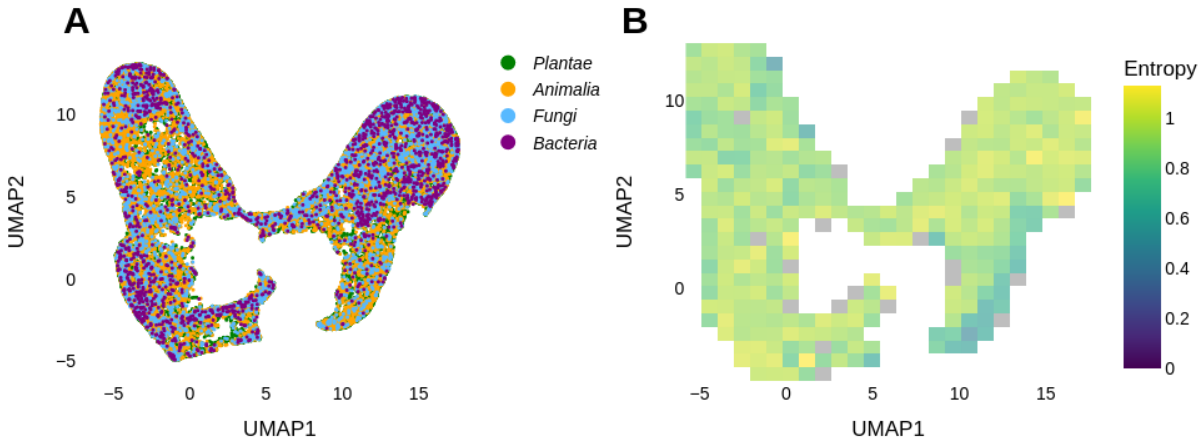


Fig 10. UMAP projection colored by different kingdoms of life.

(A) All pockets in their respective UMAP embedding space, (B) Binned entropy, S (see Materials and Methods), heatmap plot, capturing for every square bin the kingdom diversity with low S values indicating low diversity and high values corresponding to respectively increased diversity. All grid cells containing less than 20 pockets are colored gray, as no robust entropy estimate can be computed.

As we wished to discern kingdom- and species-specific differences between the pocketomes, we colored the UMAP projection by kingdom of life and species. The kingdom-specific projection map shown in Figure 10 does not reveal any strong kingdom-specific differences for *bacteria*, *fungi*, *plantae*, and *animalia*, with pockets from all kingdoms occupying all of the UMAP space (Fig 10A) and the binned entropy heatmap capturing diversity revealing a rather uniform kindgom representation. The only region that contain less binding sites found on bacterial proteins, is the lower right part of the UMAP plot (Fig 10A and 10B). This corresponds to the region where most of the pockets are found in the protein kinase domain. This eukaryotic domain is not found in any bacterial protein. Likewise, no species-specific binding sites are revealed in the respective UMAP plot shown in S10 Fig.

Taken together, these results suggest that either compound binding is not very specifically associated with pocket descriptors, or, and we consider this to be more likely, the binding site definition of DeeplyTough's embeddings may be too coarse-grained to capture local similarities between different binding sites or differences, respectively. The binding sites are rather separated

by their overall physicochemical features and high-level domain information. This leads to an embedding space that is populated too densely to allow for any meaningful fine-grained clustering.

Discussion

In this study, we attempted to generate a global view of the pocketome universe. Inspecting the 11 species for which the AlphaFold database contains a comprehensive set of predicted protein structures and applying established binding site (pocket) prediction algorithms, we performed a systematic analysis of binding site properties, ligand sets, clustering, and their scaling with proteome sizes.

Our analysis suggests that during the course of evolution, with species evolving larger proteomes and more differentiated molecular functions, unlike perhaps expected, this increase in proteome size and complexity does not commensurate with a proportional increase of binding site “innovation”. Instead, less than proportionally expected unique binding sites seem to have emerged in evolution (Fig 5A). This implies that rather than compound binding sites, other functions may have been emerged and expanded, such as possibly transcriptional and translational regulation and signaling. However, it needs to be cautioned that this conclusion is based on the approach pursued here with all the caveats pertaining to the used tools (detailed below). Furthermore, we noted that this conclusion is driven by the relatively high number of singleton binding sites, i.e. sites that are unique and are not similar to any other site in the respective species’ proteome. When inspecting binding sites with multiple instances in the respective proteome, i.e. communities, the above stated conclusion is less convincing, in fact, it seems to support a proportional relationship of binding sites and proteome size (Fig 5B). Thus, the exact nature of the scaling law awaits further scrutiny.

While the pocketome universe exhibits partitioning into pockets of particular physicochemical properties (Fig 7) and compound call binding (Fig 8), we did not identify any such segregation with regard to species or kingdom (Fig 10 A, B), except for the missing protein kinase domain in bacteria. This may either be the result of technical limitations to discern any

present species/kingdom-specific pockets (discussed below), or may suggest that indeed, species/kingdoms are not characterized by specific binding sites. However, as, for example, plants do contain specialized metabolic reactions (secondary metabolism), it seems surprising that this should be the case. Yet, as those specialized compounds are consumed and degraded by other species as food, perhaps it is a wrong assumption to expect unique pockets in plants, as in species that feed on plants similar binding sites need to exist for the processing of the compounds.

One source of uncertainty possibly affecting our analyses concerns the reliability of predicted pockets. As P2Rank [6] does not perform any sub-sampling, the predicted pockets are, on average, larger than known pockets (Fig 1B) and are likely to merge several distinct pockets, i.e., for pockets that are found in protein kinase domains. This leads to a threshold-based exclusion of large ATP binding sites that are identified with a high probability, but are predicted to comprise more than 31 residues, and therefore considered too large to be included (we introduced a size threshold of 31 residues), which may partially explain the large fraction of nucleoside/nucleotide/nucleobase pockets that are not found for some species. Introducing a sub-sampling algorithm as a post-processing step of P2Rank, could improve the precision of the binding site comparison as well as the quality of the P2Rank predictions in general.

One option to potentially refine predicted binding sites may be to exploit domain information. In our study we identified a subset of ~27,500 pockets that are found at a domain-no-domain interface. Those pockets may be less well defined structurally than pockets fully embedded in specific domains of known function and may, hence, require a focused analysis.

Inorganic ions are the second largest fraction of known pockets that are not correctly identified via P2Rank (Fig 2). P2Rank may often fail to correctly identify such sites as they tend to involve a smaller number of residues during binding. P2Rank prediction might be influenced by missing ions or cofactors as it depends on physicochemical properties. Employing AlphaFill to implant the most likely ions or cofactors into the predicted structures could potentially replace binding site prediction for this compound class [45].

Another type of ligand for which P2Rank struggles to identify the binding sites is carbohydrates, likely due to their complexity in shape, flexibility, and size. This problem has been

recently addressed by the development of several specialized prediction tools, such as DeepGlycanSite [46] and PeSTo-Carbs [47]. Specialized binding site prediction tools, such as PocketMiner [48], have also been developed for the detection of cryptic pockets. Cryptic pockets are pockets that are only apparent following a conformational change or upon ligand binding. Complementing the current data set of predicted pockets with binding sites predicted by specialized tools, is a promising route towards investigating the full space of protein binding sites.

Our study yet again reveals that binding site comparison remains a challenging task, although it has been a subject of research and tool development for many years. While in proteins there are secondary structural elements that allow for a relatively fast and relatively well-defined comparison between proteins, the comparison of protein binding sites is more complex, as they do not have a continuous sequence. For a meaningful spatial comparison, rotational and translational invariance is required for binding site comparison tools, such as implemented by DeeplyTough [17]. DeeplyTough compares binding sites in an alignment-free manner by reducing the representation of each pocket to a feature vector. While this captures the overall shape, domain information and physicochemical features of a pocket and allows for fast comparison via Euclidean distance, it is limited in the detection of similar sub-shapes within two binding pockets. Although DeeplyTough was trained to separate dissimilar binding sites by a certain margin with a contrastive loss, the embedding space of DeeplyTough is densely packed. Thus, making it difficult to obtain a clustering for inter-species comparison as true positives (= similar pockets) as well as false positives (= dissimilar binding sites) are close in the Euclidean space. One reason for that might be pocket types encountered here, but that were not found in the limited training data set.

In contrast to alignment-free methods, alignment-based methods, such as ProBiS [27] allow insights into which amino acids residues or surface atoms of the compared pockets are similar and contribute positively to the comparison score. This allows to easily interpret why two binding sites are considered similar or different and, importantly, allows for a sensitive identification of local sub-structural matches. However, these approaches are computationally too expensive at the moment to allow large-scale, inter-species studies on tens of thousands of binding sites as attempted here [26]. Hence, the development of a fast and reliable alignment-

based or alignment-free comparison method is still needed to improve inter-species binding site comparison and clustering at scale.

On the other hand, the search space of potentially similar binding sites could be reduced by additional filtering to make alignment-based methods feasible for an all-against-all inter-species comparison. Binding sites that are known to be dissimilar could be pre-filtered by their amino acid composition or their overall physicochemical properties. Binding sites that are known to be similar could be pre-clustered based on the domain they belong to. This would reduce the number of pairwise comparisons needed especially for redundant domains, such as the protein kinase domain.

Many of the recently published binding site comparison tools have been evaluated on the ProSPECCTs datasets [25]. ProSPECCTs is an ensemble of seven different data sets of pairs of binding sites with known similarities. In contrast to the aim of this study, the ProSPECCTs datasets do not contain all-against-all comparisons of the respective binding sites. Therefore, the actual performance of the binding sites comparison tools might be overestimated. In addition, the datasets were derived from known protein-ligand complexes. Reim et al. [26] investigated the performance of their tool, SiteMine, for predicted pockets and observed that it is performing worse for some datasets. As predicted pockets consist, on average, of more residues, they might lose some specificity and overall physicochemical properties represented unspecifically or even incorrectly. This might lead to pockets that are more similar than they are in reality. Hence, complicating the binding sites comparison at a large scale. DeeplyTough might lose additional specificity because it is using all surrounding residues of the center of mass of the pocket for embedding and the usage of predicted pockets instead of known pockets as they contain on average more residues than known pockets.

In addition, the use of predicted, monomeric protein structures from AlphaFold database results in different challenges that need to be addressed in future research. These challenges include protein-protein interactions, post-translational modifications and conformational changes that occur upon binding. Currently, the AlphaFold database only includes monomeric structures. However, studies for, e.g., HUMAN and YEAST, suggest that most proteins are at least part of one protein complex [49,50]. Structural predictions might differ between monomer and multimer

predictions. Therefore, it would be desirable to investigate potential binding sites from a multimer perspective. Although multimer prediction is challenging due to often unknown stoichiometries and the exploding combinatorics with higher number of chains, reliable prediction of structures for dimers has been achieved by new models, such as AlphaFold-Multimer [51]. The challenge to identify potential interaction partner(s) for each protein of interest through, e.g., protein-protein interaction prediction, remains.

The AF database contains one predicted structure per protein. However, there are proteins that change their conformation upon binding or have an ensemble of different conformations. These conformational changes are currently not taken into consideration. In addition, studies showed that AF tends to predict the “mean” structure in cases where apo and holo structure of a protein differ [52]. This might influence the prediction and classification of so-called cryptic pockets that are only found in holo-structures. This problem can be tackled by established methods, such as normal mode analysis (NMA), as well as DL-based methods that have been developed recently to overcome this issue [53,54].

In addition, the AF database does not consider any post-translational modifications which might influence the structural conformation of proteins. Prominently, protein phosphorylation has been discussed as a modulator of compound binding [55]. A recent, PDB-wide study [56] could show that for phosphorylation these conformational changes are rather small with only ~30% of phosphorylation events causing a RMSD ≥ 2 Å as compared to the non-phosphorylated structure. However, this number might be too conservative as the PDB database is biased towards less flexible conformations as conformational flexible structures are harder to crystallize. In the same study, the authors could identify a small subset of structures for which phosphorylation is mechanically coupled to functional sites. Extending their work to AF predicted structures, might be a promising avenue of future research.

Another caveat concerns the notion of binding specificity. We approached our study by positing that novel ligands require novel binding sites. However, it is known that binding sites as well as ligands can be promiscuous [14], thereby a one-to-one relationship will not always be true. As a switch from selective to promiscuous binding may involve only small structural changes, functional diversity may grow even in the absence of easily noticeable structural changes.

Lastly, it is known that a substantial portion of proteins and regions within proteins are structurally disordered under physiological conditions [57]. However, it is also known or surmised that these proteins or protein regions do engage in (specific) molecular interactions, including binding to small molecules [58,59]. While structure-based binding site detection in disordered regions are challenging, attempts have been made to apply molecular dynamics simulation to interactions of compounds with disordered protein regions [60], or to predict sequence regions that are likely to interact with other molecules, for example, proteins [61]. Future computational predictions of small molecules with disordered regions of proteins will critically depend on the availability of large-enough experimental datasets as generated by specific NMR protocols [62] and others.

Although some of the challenges mentioned above, such as the post-translational modification, are addressed by AlphaFold3 [63], structures produced by AF3 are not yet available at such a large scale as they can be found in the AF database. Hence, they cannot yet be used for a proteome-wide comparison of binding pockets.

Conclusions

To our knowledge, we present the first large-scale intra- and inter-species study of compound binding sites on protein surfaces using AlphaFold structures to arrive at a map of the pocketome universe. The binding site comparisons conducted for different species using ProBis reveals that more species with larger and more diverse proteomes might have relatively less diverse pocketomes as compared to species with smaller proteomes. Applying DeeplyTough for inter-species binding site comparison, we identified shortcomings of embedding-based binding site comparison tools that make them unfeasible for large-scale, inter-species clustering of binding sites. An improved methodology is needed to fully discern the scaling laws of species-specific pocketomes and to allow for inter-species clustering of predicted binding sites of the whole AlphaFold database.

Acknowledgements

We wish to thank Zoran Nikoloski, Ariane Nunes Alves, and Alisdair Fernie for helpful discussions.

Data availability

Interactive Plotly plots for the UMAP projections of the DeeplyTough embeddings, which allow filtering for specific compound classes or species, are available as HTML file via Zenodo (https://doi.org/10.5281/zenodo.14800168).

References

1. Eguida M, Rognan D. Estimating the Similarity between Protein Pockets. Int J Mol Sci. 2022;23. doi:10.3390/ijms232012462

2. Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: An open source platform for ligand pocket detection. BMC Bioinformatics. 2009;10: 1–11. doi:10.1186/1471-2105-10-168

3. Desaphy J, Azdimousa K, Kellenberger E, Rognan D. Comparison and druggability prediction of protein-ligand binding sites from pharmacophore-annotated cavity shapes. J Chem Inf Model. 2012;52: 2287–2299. doi:10.1021/ci300184x

4. An J, Totrov M, Abagyan R. Pocketome via Comprehensive Identification and Classification of Ligand Binding Envelopes. Mol Cell Proteomics. 2005;4: 752–761. doi:10.1074/MCP.M400159-MCP200

5. Halgren TA. Identifying and characterizing binding sites and assessing druggability. J Chem Inf Model. 2009;49: 377–389. doi:10.1021/ci800324m

6. Krivák R, Hoksza D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. J Cheminformatics. 2018;10: 39. doi:10.1186/s13321-018-0285-8

7. Volkamer A, Eid S, Turk S, Jaeger S, Rippmann F, Fulle S. Pocketome of human kinases: Prioritizing the ATP binding sites of (Yet) untapped protein kinases for drug discovery. J Chem Inf Model. 2015;55: 538–549. doi:10.1021/ci500624s

8. Hedderich JB, Persechino M, Becker K, Heydenreich FM, Gutermuth T, Bouvier M, et al. The pocketome of G-protein-coupled receptors reveals previously untargeted allosteric sites. Nat Commun 2022 131. 2022;13: 1–12. doi:10.1038/s41467-022-29609-6

9. Kufareva I, Ilatovskiy AV, Abagyan R. Pocketome: an encyclopedia of small-molecule binding sites in 4D. Nucleic Acids Res. 2012;40: D535. doi:10.1093/NAR/GKR825

10. Bhagavat R, Sankar S, Srinivasan N, Chandra N. An Augmented Pocketome: Detection and Analysis of Small-Molecule Binding Pockets in Proteins of Known 3D Structure. Structure. 2018;26: 499-512.e2. doi:10.1016/j.str.2018.02.001

11. Anand P, Chandra N. Characterizing the pocketome of Mycobacterium tuberculosis and application in rationalizing polypharmacological target selection. Sci Rep 2014 41. 2014;4: 1–17. doi:10.1038/srep06356

12. Konc J, Janežič D. ProBiS-Fold Approach for Annotation of Human Structures from the AlphaFold Database with No Corresponding Structure in the PDB to Discover New Druggable Binding Sites. J Chem Inf Model. 2022;62: 5821–5829. doi:10.1021/acs.jcim.2c00947

13. Pallante L, Cannariato M, Androutsos L, Zizzi EA, Bompotas A, Hada X, et al. VirtuousPocketome: a computational tool for screening protein–ligand complexes to identify similar binding sites. Sci Rep 2024 141. 2024;14: 1–14. doi:10.1038/s41598-024-56893-7

14. Korkuć P, Walther D. Physicochemical characteristics of structurally determined metabolite-protein and drug-protein binding events with respect to binding specificity. Front Mol Biosci. 2015;2. doi:10.3389/fmolb.2015.00051

15. Walther D. Specifics of Metabolite-Protein Interactions and Their Computational Analysis and Prediction. In: Skirycz A, Luzarowski M, Ewald JC, editors. Cell-Wide Identification of Metabolite-Protein Interactions. New York, NY: Springer US; 2023. pp. 179–197. doi:10.1007/978-1-0716-2624-5_12

16. Binkowski TA, Joachimiak A. Protein functional surfaces: Global shape matching and local spatial alignments of ligand binding sites. BMC Struct Biol. 2008;8: 1–23. doi:10.1186/1472-6807-8-45

17. Simonovsky M, Meyers J. DeeplyTough: Learning Structural Comparison of Protein Binding Sites. J Chem Inf Model. 2020;60: 2356–2366. doi:10.1021/acs.jcim.9b00554

18. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2022;50: D439–D444. doi:10.1093/NAR/GKAB1061

19. Dobson L, Szekeres LI, Gerdán C, Langó T, Zeke A, Tusnády GE. TmAlphaFold database: membrane localization and evaluation of AlphaFold2 predicted alpha-helical transmembrane protein structures. Nucleic Acids Res. 2023;51: D517–D522. doi:10.1093/nar/gkac928

20. Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. Nat Commun. 2018;9: 2542. doi:10.1038/s41467-018-04964-5

21. Bateman A, Martin MJ, Orchard S, Magrane M, Ahmad S, Alpi E, et al. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023;51: D523–D531. doi:10.1093/NAR/GKAC1052

22. Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, et al. InterPro in 2022. Nucleic Acids Res. 2023;51: D418–D427. doi:10.1093/nar/gkac993

23. Barrio-Hernandez I, Yeo J, Jänes J, Mirdita M, Gilchrist CLM, Wein T, et al. Clustering predicted structures at the scale of the known protein universe. Nature. 2023;622: 637–645. doi:10.1038/s41586-023-06510-w

24. Hastings J, De Matos P, Dekker A, Ennis M, Harsha B, Kale N, et al. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic Acids Res. 2012;41: D456–D463. doi:10.1093/nar/gks1146

25. Ehrt C, Brinkjost T, Koch O. A benchmark driven guide to binding site comparison: An exhaustive evaluation using tailor-made data sets (ProSPECCTs). PLOS Comput Biol. 2018;14: e1006483. doi:10.1371/JOURNAL.PCBI.1006483

26. Reim T, Ehrt C, Graef J, Günther S, Meents A, Rarey M. SiteMine: Large-scale binding site similarity searching in protein structure databases. Arch Pharm (Weinheim). 2024;357: 2300661. doi:10.1002/ARDP.202300661

27. Konc J, Janežič D. ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. Bioinformatics. 2010;26: 1160–1168. doi:10.1093/BIOINFORMATICS/BTQ100

28. Da Silva F, Desaphy J, Rognan D. IChem: A Versatile Toolkit for Detecting, Comparing, and Predicting Protein–Ligand Interactions. ChemMedChem. 2018;13: 507–510. doi:10.1002/cmdc.201700505

29. Batista J, Hawkins PC, Tolbert R, Geballe MT. SiteHopper - a unique tool for binding site comparison. J Cheminformatics. 2014;6: P57, 1758-2946-6-S1-P57. doi:10.1186/1758-2946-6-S1-P57

30. Ren J, Xie L, Li WW, Bourne PE. SMAP-WS: a parallel web service for structural proteome-wide ligand-binding site comparison. Nucleic Acids Res. 2010;38: W441–W444. doi:10.1093/nar/gkq400

31. Wood DJ, Vlieg JD, Wagener M, Ritschel T. Pharmacophore Fingerprint-Based Approach to Binding Site Subpocket Similarity and Its Application to Bioisostere Replacement. J Chem Inf Model. 2012;52: 2031–2043. doi:10.1021/ci3000776

32. Pedregosa F, Michel V, Grisel O, Blondel M, Prettenhofer P, Weiss R, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12: 2825–2830.

33. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv; 2020. Available: http://arxiv.org/abs/1802.03426

34. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25: 1422–1423. doi:10.1093/BIOINFORMATICS/BTP163

35. Lobry JR, Gautier C. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes. Nucleic Acids Res. 1994;22: 3174–3180. doi:10.1093/nar/22.15.3174

36. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982;157: 105–132. doi:10.1016/0022-2836(82)90515-0

37. Shrake A, Rupley JA. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. J Mol Biol. 1973;79: 351–371. doi:10.1016/0022-2836(73)90011-9

38. Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. Maximum Allowed Solvent Accessibilites of Residues in Proteins. Porollo A, editor. PLoS ONE. 2013;8: e80635. doi:10.1371/journal.pone.0080635

39. Hagberg AA, Schult DA, Swart PJ. Exploring Network Structure, Dynamics, and Function using NetworkX. Pasadena, California; 2008. pp. 11–15. doi:10.25080/TCWV9851

40. Traag V, Waltman L, Eck NJ van. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep. 2019;9: 5233. doi:10.1038/s41598-019-41695-z

41. Rossetti G, Milli L, Cazabet R. CDLIB: a python library to extract, compare and evaluate communities from complex networks. Appl Netw Sci. 2019;4: 52. doi:10.1007/s41109-019-0165-9

42. Meng EC, Goddard TD, Pettersen EF, Couch GS, Pearson ZJ, Morris JH, et al. UCSF CHIMERAX : Tools for structure building and analysis. Protein Sci. 2023;32: e4792. doi:10.1002/pro.4792

43. Collaborative data science. Montreal, QC: Plotly Technologies Inc.; 2015. Available: https://plot.ly

44. Koonin EV, Wolf YI, Karev GP. Power Laws, Scale-Free Networks and Genome Biology. Boston, MA: Springer US; 2006. doi:10.1007/0-387-33916-7

45. Hekkelman ML, De Vries I, Joosten RP, Perrakis A. Nature Methods nature methods AlphaFill: enriching AlphaFold models with ligands and cofactors. 2023. doi:10.1038/s41592-022-01685-y

46. He X, Zhao L, Tian Y, Li R, Chu Q, Gu Z, et al. Highly accurate carbohydrate-binding site prediction with DeepGlycanSite. Nat Commun. 2024;15: 5163. doi:10.1038/s41467-024-49516-2

47. Bibekar P, Krapp L, Peraro MD. PeSTo-Carbs: Geometric Deep Learning for Prediction of Protein–Carbohydrate Binding Interfaces. J Chem Theory Comput. 2024;20: 2985–2991. doi:10.1021/acs.jctc.3c01145

48. Meller A, Ward M, Borowsky J, Kshirsagar M, Lotthammer JM, Oviedo F, et al. Predicting locations of cryptic pockets from single protein structures using the PocketMiner graph neural network. Nat Commun. 2023;14: 1177. doi:10.1038/s41467-023-36699-3

49. Gavin A-C, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature. 2002;415: 141–147. doi:10.1038/415141a

50. Huttlin EL, Bruckner RJ, Paulo JA, Cannon JR, Ting L, Baltier K, et al. Architecture of the human interactome defines protein communities and disease networks. Nature. 2017;545: 505–509. doi:10.1038/nature22366

51. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, et al. Protein complex prediction with AlphaFold-Multimer. Bioinformatics; 2021. doi:10.1101/2021.10.04.463034

52. Saldaño T, Escobedo N, Marchetti J, Zea DJ, Mac Donagh J, Velez Rueda AJ, et al. Impact of protein conformational diversity on AlphaFold predictions. Valencia A, editor. Bioinformatics. 2022;38: 2742–2748. doi:10.1093/bioinformatics/btac202

53. Wayment-Steele HK, Ojoawo A, Otten R, Apitz JM, Pitsawong W, Hömberger M, et al. Predicting multiple conformations via sequence clustering and AlphaFold2. Nat 2023 6257996. 2023;625: 832–839. doi:10.1038/s41586-023-06832-9

54. Monteiro Da Silva G, Cui JY, Dalgarno DC, Lisi GP, Rubenstein BM. High-throughput prediction of protein conformational distributions with subsampled AlphaFold2. Nat Commun. 2024;15: 2464. doi:10.1038/s41467-024-46715-9

55. Korkuć P, Walther D. Spatial proximity statistics suggest a regulatory role of protein phosphorylation on compound binding. Proteins Struct Funct Bioinforma. 2016;84: 565–579. doi:10.1002/prot.25001

56. Correa Marrero M, Mello VH, Sartori P, Beltrao P. Global comparative structural analysis of responses to protein phosphorylation. Bioinformatics; 2024. doi:10.1101/2024.10.18.617420

57. Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. Nat Rev Mol Cell Biol. 2015;16: 18–29. doi:10.1038/nrm3920

58. Uversky VN. Intrinsic Disorder-based Protein Interactions and their Modulators. Curr Pharm Des. 2013;19: 4191–4213. doi:10.2174/1381612811319230005

59. Mollica L, Bessa LM, Hanoulle X, Jensen MR, Blackledge M, Schneider R. Binding Mechanisms of Intrinsically Disordered Proteins: Theory, Simulation, and Experiment. Front Mol Biosci. 2016;3. doi:10.3389/fmolb.2016.00052

60. Herrera-Nieto P, Pérez A, De Fabritiis G. Small Molecule Modulation of Intrinsically Disordered Proteins Using Molecular Dynamics Simulations. J Chem Inf Model. 2020;60: 5003–5010. doi:10.1021/acs.jcim.0c00381

61. Mészáros B, Simon I, Dosztányi Z. Prediction of Protein Binding Regions in Disordered Proteins. Casadio R, editor. PLoS Comput Biol. 2009;5: e1000376. doi:10.1371/journal.pcbi.1000376

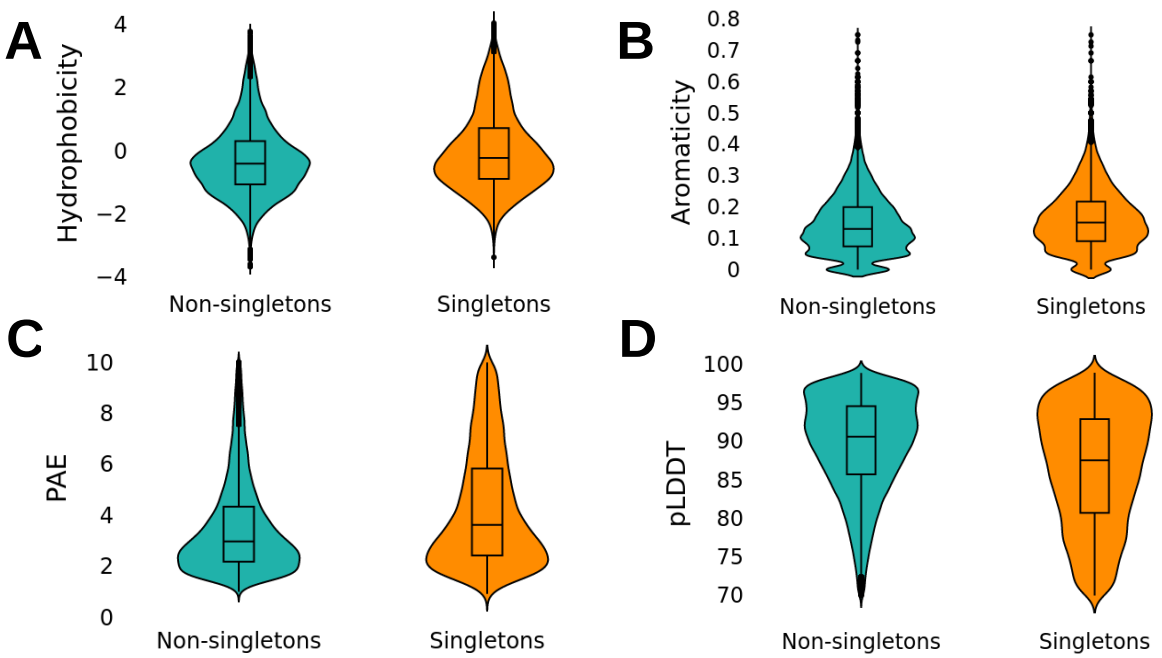
62. Brutscher B, Felli IC, Gil-Caballero S, Hošek T, Kümmerle R, Piai A, et al. NMR Methods for the Study of Instrinsically Disordered Proteins Structure, Dynamics, and Interactions: General Overview and Practical Guidelines. In: Felli IC, Pierattelli R, editors. Intrinsically Disordered Proteins Studied by NMR Spectroscopy. Cham: Springer International Publishing; 2015. pp. 49–122. doi:10.1007/978-3-319-20164-1_3

63. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nat 2024. 2024; 1–3. doi:10.1038/s41586-024-07487-w

Supporting material

S1 Table. Additional information for all predicted pockets including the ratio of pockets per protein, the number of pockets before filtering and the number of pockets that did not pass the respective quality check.

Org.	Gene dupl. [%]	# pockets before filtering	# pLDDT < 70	# PAE > 10	# TM checks passed < 6	# pockets / # proteins
ECOLI	0.9 %	1,754	58	62	16	0.437
YEAST	2.2 %	2,697	274	302	51	0.409
CANAL	12.6 %	2,652	209	259	74	0.405
ARATH	34.7 %	10,006	876	1,010	258	0.345
MAIZE	53.3 %	13,873	2,703	2,615	690	0.288
ORYSJ	6.3 %	11,273	1,722	1,828	340	0.26
SOYBN	73.6 %	20,172	2,531	2,586	762	0.338
DROME	58.2 %	4,913	313	440	123	0.346
CAEEL	25.7 %	7,561	1,019	1,016	202	0.345
MOUSE	48.0 %	8,725	573	774	125	0.374
HUMAN	62.4 %	7,873	553	947	172	0.354



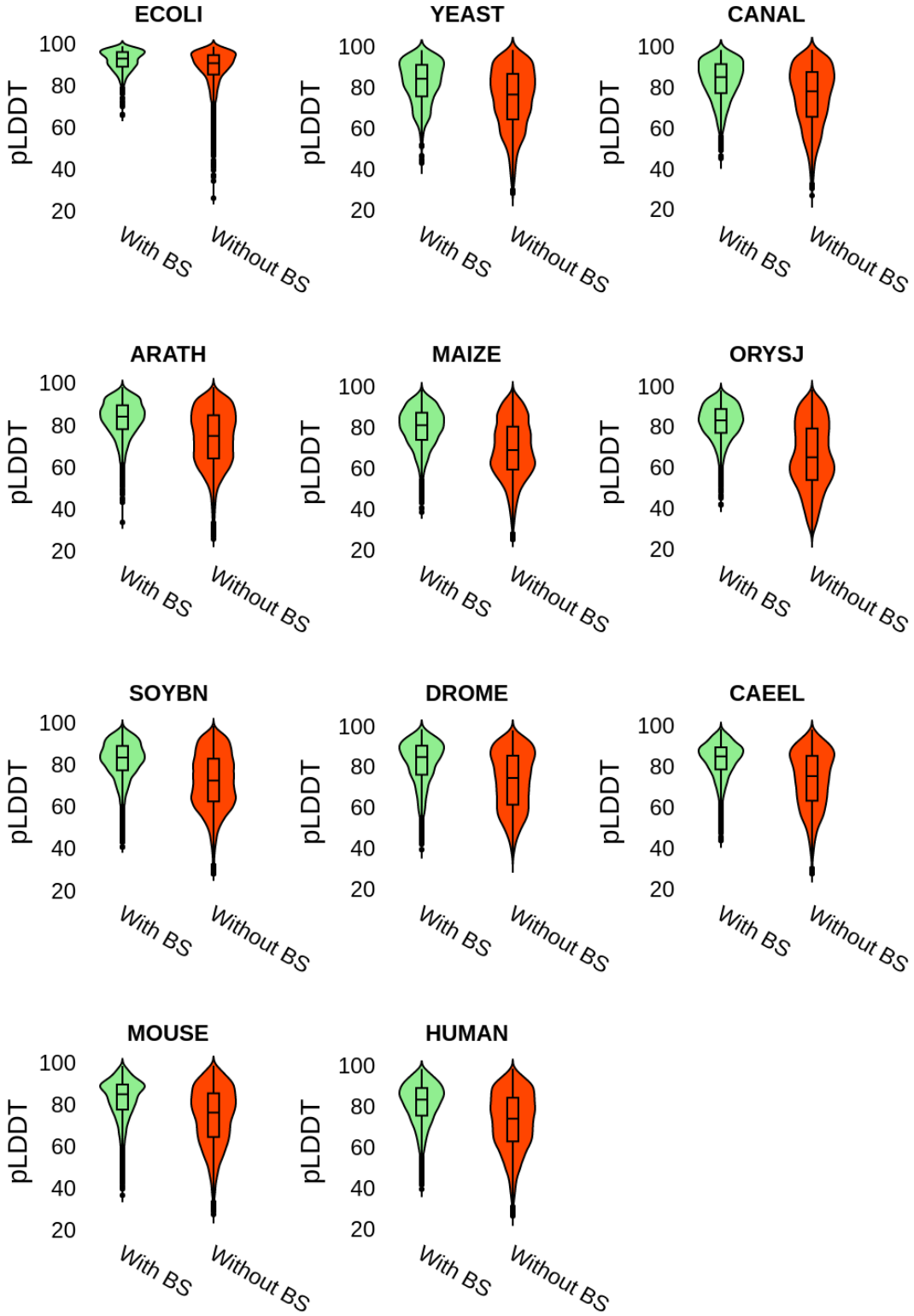
S1 Fig Distribution of different properties plotted for non-singletons and singletons. Although all distributions show significant differences between the two groups (p-value < 0.01), the respective effect sizes (Cohen's d) can be small, as is the case for hydrophobicity, aromaticity, PAE, and pLDDT. The largest difference for these properties was

observed for the pLDDT with an effect size of 0.337. For hydrophobicity, aromaticity, and PAE, the effect size between non-singletons and singletons is 0.091, 0.024 and 0.27, respectively.

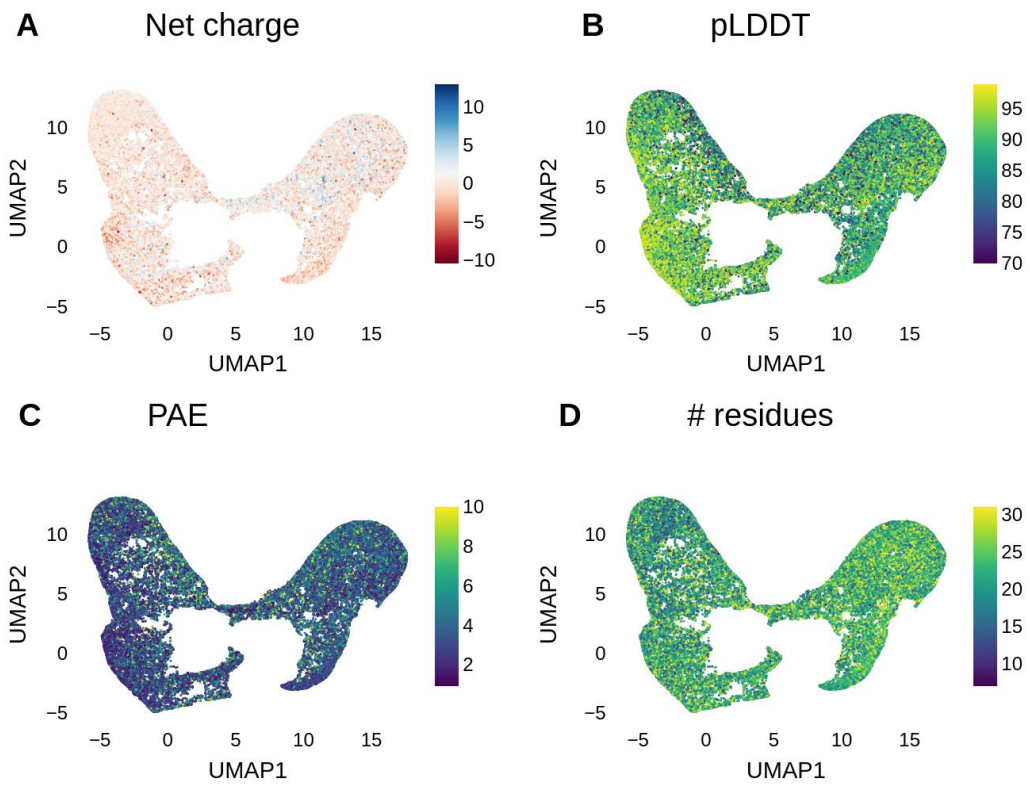
S2 Table. Number of unique pockets, singletons, and real clusters with threshold 5 applied to the ProBiS alignment score.

Species	# Unique BS	# Singletons	# Real clusters	# FS cluster
ECOLI	1,208	1,028	180	3,021
YEAST	1,527	1,311	216	4,775
CANAL	1,578	1,378	200	5,026
ARATH	3,040	2,142	898	14,507
MAIZE	3,488	2,437	1,051	21,417
ORYSJ	3,002	2,232	770	21,669
SOYBN	4,069	2,453	1,616	20,801
DROME	2,562	2,180	382	9,794
CAEEL	3,081	2,551	530	11,955
MOUSE	3,219	2,503	716	13,137
HUMAN	3,266	2,584	682	13,546

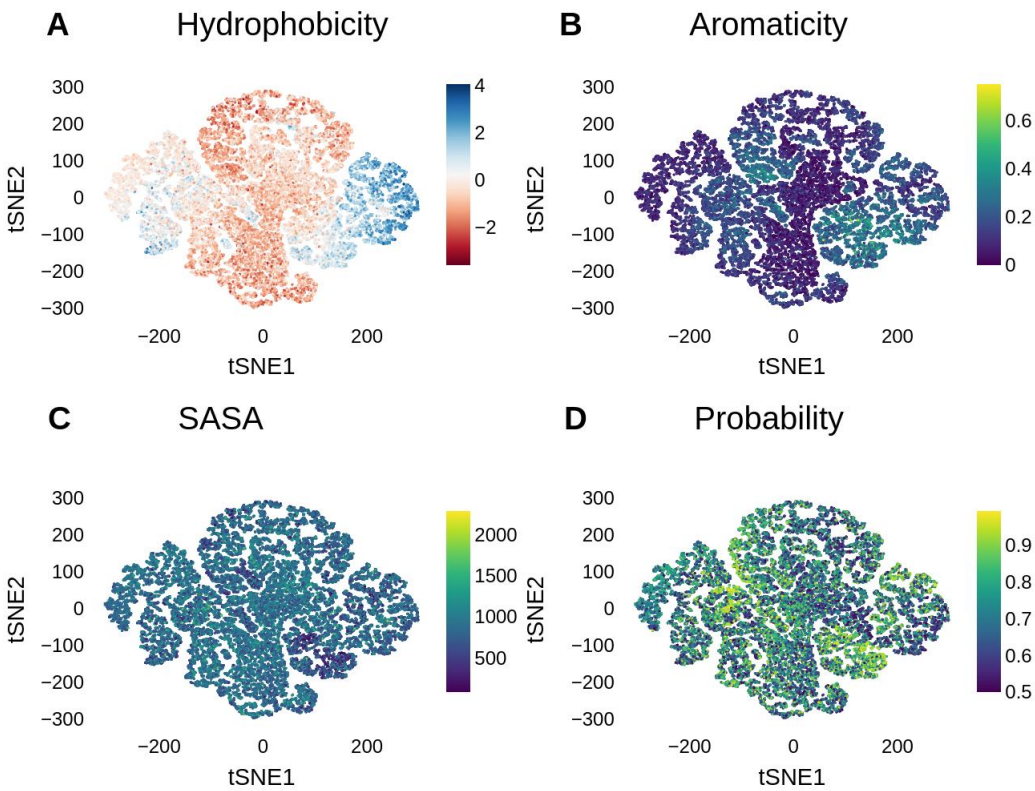
The threshold was applied to the similarity matrix that was obtained using the alignment scores of ProBiS.



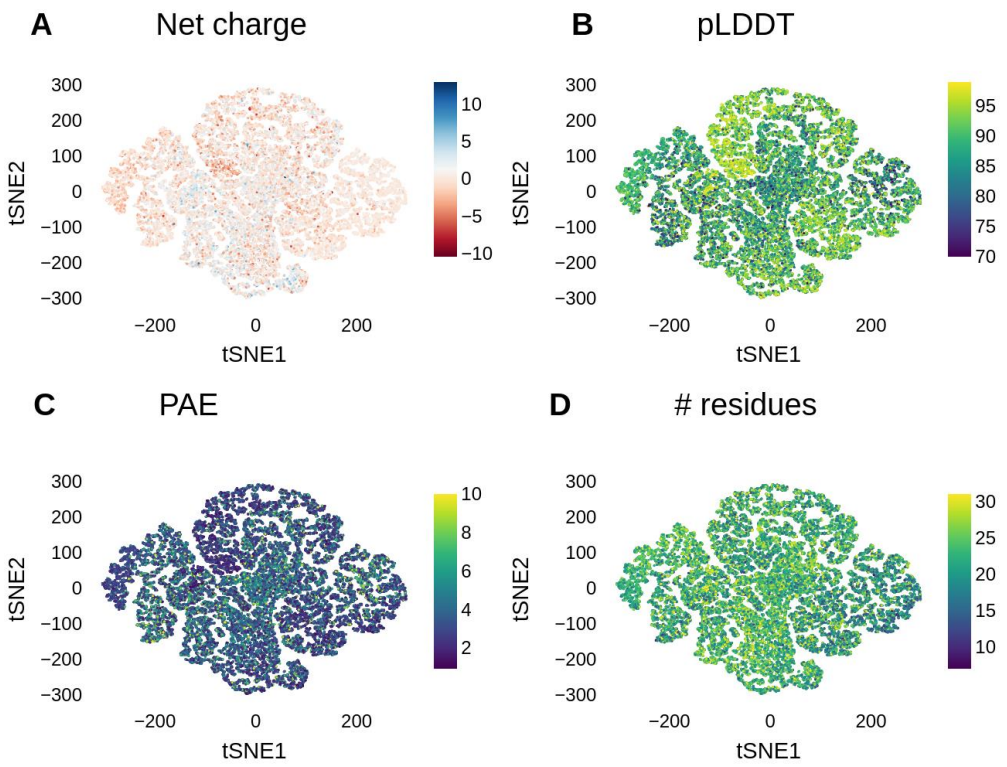
S2 Fig Violin plots for all species used in the study for the pLDDT of proteins with and without binding sites. The highest median pLDDT for proteins with and without binding sites was observed for ECOLI. The lowest median pLDDT of proteins with binding sites was observed for MAIZE with 81.05 and for ORYSJ of proteins without binding sites with 67.24. The median pLDDT values of proteins without binding sites for the other species range from 68.85 to 91.41.



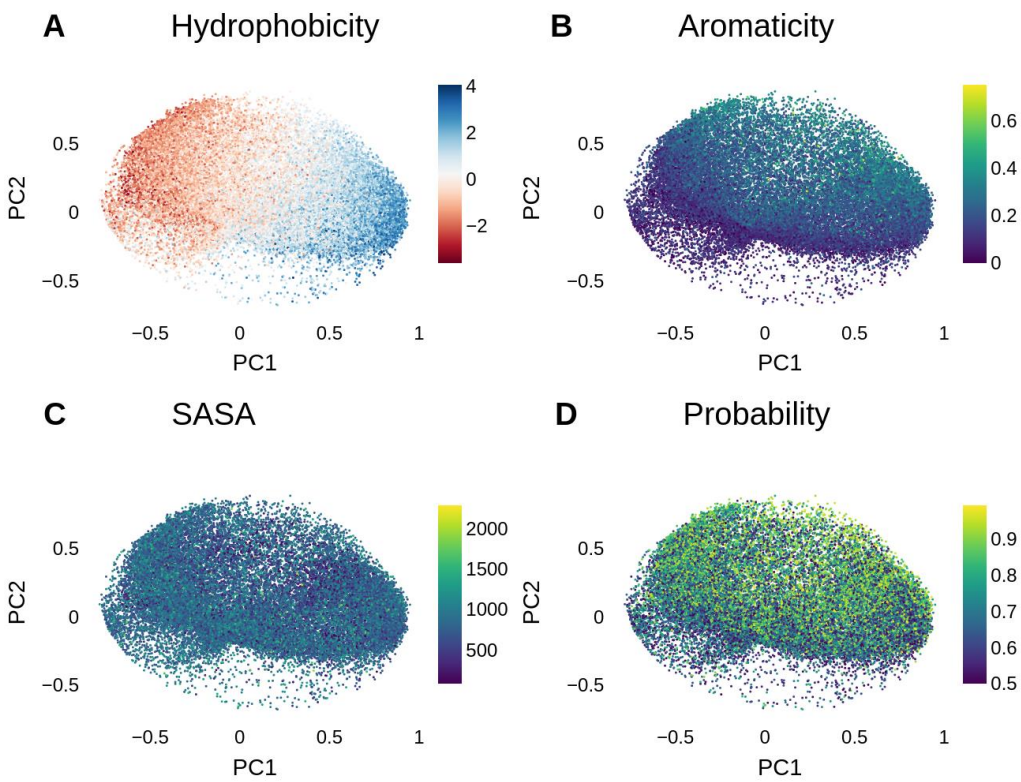
S3 Fig UMAP plots for all predicted pockets colored by different properties.
The first component of the UMAP projection reflects to some extent net charge, pLDDT, PAE, and # (number of) residues lining the pocket. The second component reflects to some extent pLDDT and PAE. However, the correlation for all the displayed properties (to the first two projection dimensions) is lower than for aromaticity and hydrophobicity.



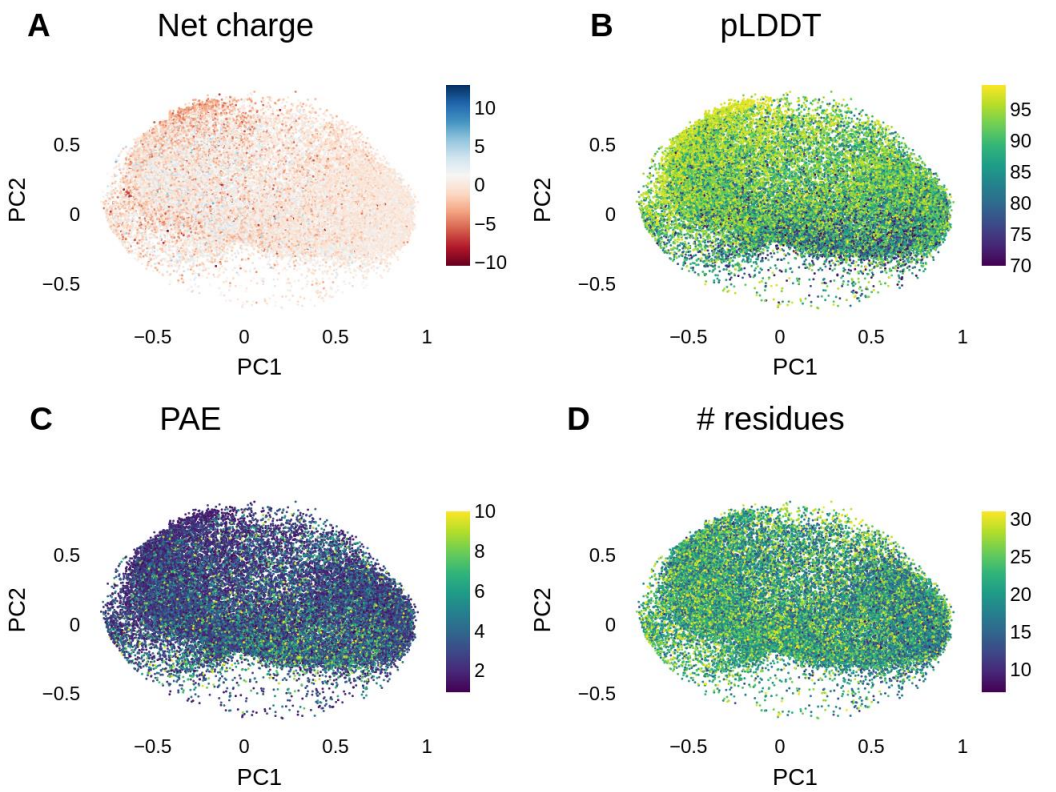
S4 Fig tSNE plots for all predicted pockets colored by different properties.
DeeplyTough seems to be able to differentiate pockets based on their hydrophobicity and aromaticity. While other properties such as SASA and probability seem to be less important. The latter is expected as both properties are no explicit input feature of DeeplyTough.



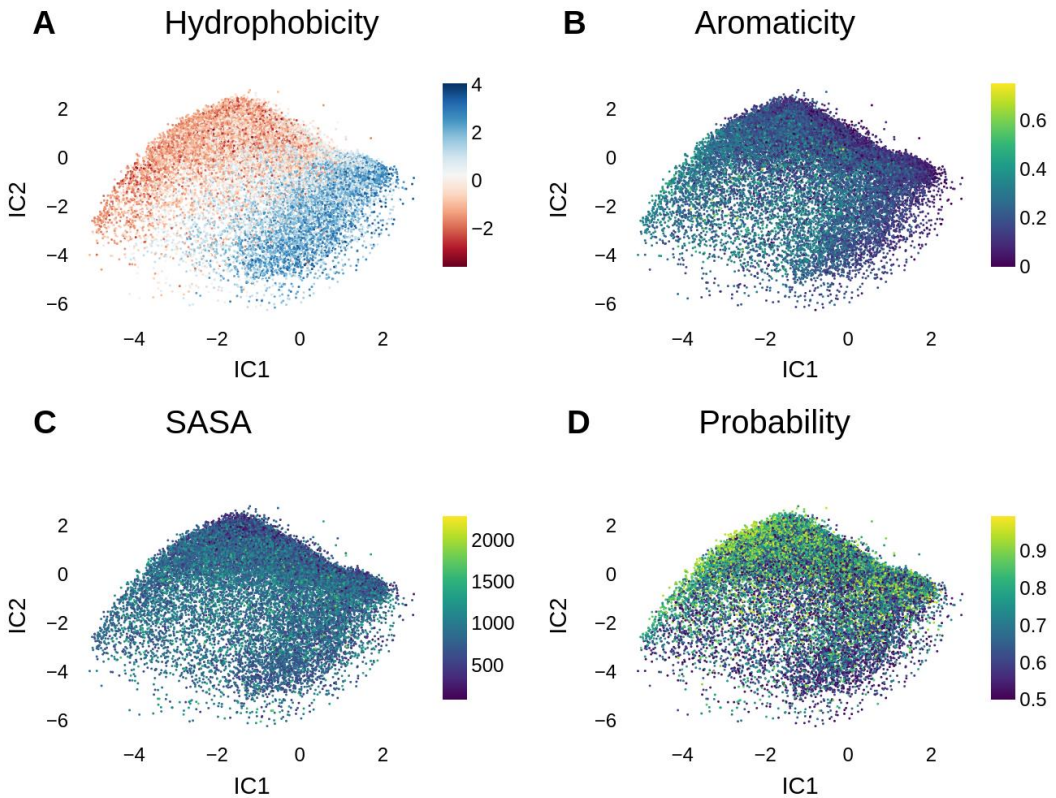
S5 Fig tSNE plots for all predicted pockets colored by different properties.
While the first component of the tSNE projection reflects to some extent pLDDT and PAE, the second component of the tSNE plot does not correlate with any of the properties of interest, including net charge, pLDDT, PAE, and # residues (number of residues lining the pocket).



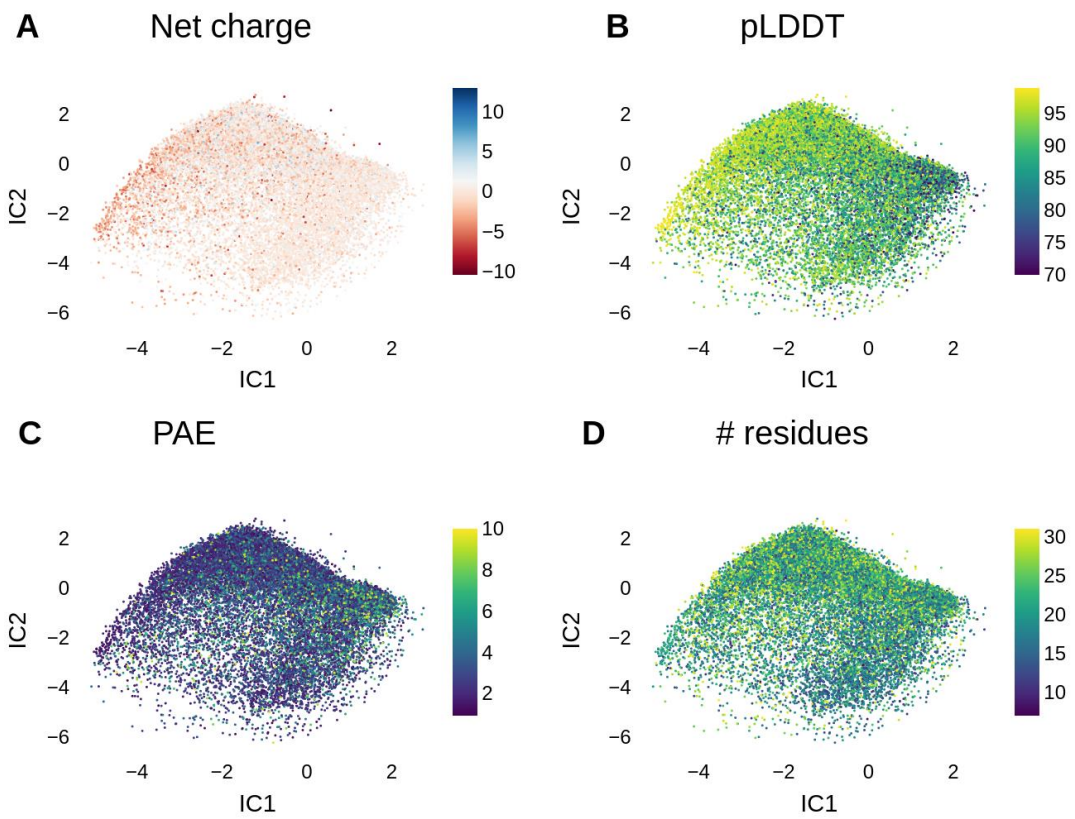
S6 Fig PCA plots for all predicted pockets colored by different properties.
DeeplyTough seems to be able to differentiate pockets based on their hydrophobocity and aromaticity. While other properties such as SASA and probability seem to be less important. The latter is expected as both properties are no explicit input features of DeeplyTough.



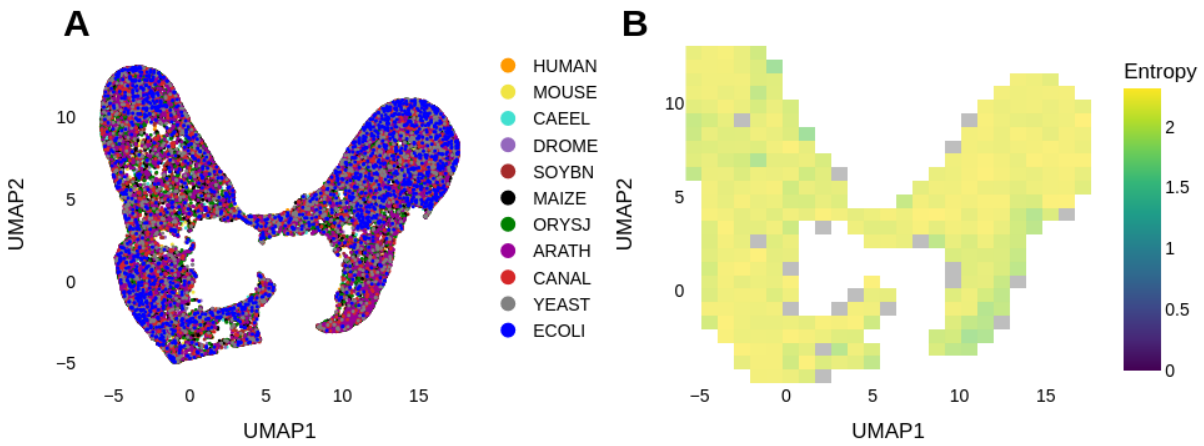
S7 Fig PCA plots for all predicted pockets colored by different properties.
The first component of the PCA projection reflects to some extent on pLDDT and # residues. The second component reflects to some extent pLDDT and PAE. However, the correlation for all the displayed properties is lower than for aromaticity and hydrophobicity.



S8 Fig ICA plots for all predicted pockets colored by different properties.
DeeplyTough seems to be able to differentiate pockets based on their hydrophobocity and aromaticity. While other properties such as SASA and probability seem to be less important. The latter is expected as both properties are no explicit input feature of DeeplyTough.



S9 Fig ICA plots for all predicted pockets colored by different properties.
While the first component of the UMAP projection reflects to some extent on the pLDDT, the second component reflects to some extent on pLDDT and PAE. However, the correlation for all the displayed properties is lower than for aromaticity and hydrophobicity.



S10 Fig UMAP plot colored by species and binned entropy.
(A) UMAP projection colored by different species. All pockets in their respective UMAP embedding space, (B) Binned entropy, S (see Materials and Methods), heatmap plot, capturing for every square bin the species diversity with low entropy, S, values indicating low diversity and high values corresponding to respectively increased diversity. All grid cells containing less than 20 pockets are colored gray.