Rapid proteome-wide prediction of lipid-interacting proteins through ligand-guided structural genomics

Jonathan Chiu-Chun Chou,[1] Cassandra M. Decosto,[1#] Poulami Chatterjee,[1#] Laura M. K. Dassama[1,2*]

[1]Department of Chemistry and Sarafan ChEM-H Institute, Stanford University, Stanford, CA 94305

[2]Department of Microbiology and Immunology, Stanford School of Medicine, Stanford, CA 94305

[#]Equal contributors

[*]Correspondence to: dassama@stanford.edu

Abstract

Lipids are primary metabolites that play essential roles in multiple cellular pathways. Alterations in lipid metabolism and transport are associated with infectious diseases and cancers. As such, proteins involved in lipid synthesis, trafficking, and modification, are targets for therapeutic intervention. The ability to rapidly detect these proteins can accelerate their biochemical and structural characterization. However, it remains challenging to identify lipid binding motifs in proteins due to a lack of conservation at the amino acid level. Therefore, new bioinformatic tools that can detect conserved features in lipid binding sites are necessary. Here, we present Structure-based Lipid-interacting Pocket Predictor (SLiPP), a structural bioinformatics algorithm that uses machine learning to detect protein cavities capable of binding to lipids in experimental and AlphaFold-predicted protein structures. SLiPP, which can be used at proteome-wide scales, predicts lipid binding pockets with an accuracy of 96.8% and a F1 score of 86.9%. Our analyses revealed that the algorithm relies on hydrophobicity-related features to distinguish lipid binding pockets from those that bind to other ligands. Use of the algorithm to detect lipid binding proteins in the proteomes of various bacteria, yeast, and human have produced hits annotated or verified as lipid binding proteins, and many other uncharacterized proteins whose functions are not discernable from sequence alone. Because of its ability to identify novel lipid binding proteins, SLiPP can spur the discovery of new lipid metabolic and trafficking pathways that can be targeted for therapeutic development.

Introduction

As the main components of cell envelopes, lipids are essential building blocks of life that provide a barrier to the cell. In higher-order organisms, the composition of lipids in membranes enclosing organelles is crucial for the identity and function of those organelles. Lipids also serve as a source of energy, and they and their derivatives function as signaling molecules. Mis-regulation of lipids leads to several human diseases, including Niemann-Pick disease[1], Farber's disease[2], Barth syndrome[3], Wolman's disease[4], and more. Furthermore, altered lipid metabolism is a hallmark of cancer[5], as increased lipid synthesis and uptake is critical for the rapid growth of cancer cells. Lipid acquisition and metabolism is also important in infectious diseases[6-8], particularly in the context of infections mediated by pathogenic bacteria that lack the machinery for de novo lipid synthesis or use host lipids during colonization. The former include spirochetes such as *Borrelia burgdorferi*[9] and *Treponema pallidum*[10], while the latter includes the intracellular pathogen *Chlamydia trachomatis*[11,12]. Despite the relevance of lipids in biology, there is a limited set of available tools for large scale identification of proteins that engage with these molecules. Whereas chemoproteomics[13,14] and gene expression[15,16] studies have been powerful for revealing lipid interactomes, they are limited to culturable systems and require highly specialized tools (e.g., modified lipids as probes and mass spectrometers). Additionally, these methods may not accurately detect lipid interacting proteins that are present in low abundance in biological samples. In theory, bioinformatics can aid in overcoming these challenges, but the principles that govern the recognition of lipids remain poorly defined.

Historically, the prediction of protein function relied on protein sequence similarity. The commonly used Basic Local Alignment Search Tool (BLAST)[17] was first developed in 1990 and infers functional homology through sequence homology. Similarly, the hidden Markov model[18], introduced in 1998, categorizes protein families to further imply the shared functions. With the emergence of machine learning and neural networks, newer models including ProteInfer[19], ProLanGO[20], DeepGO[21], and DeepGOPlus[22], have been developed. These methods all rely solely on protein sequence similarity to infer functional homology. Recently, other methods leveraging the structure prediction capabilities of AlphaFold[23] have attempted to use structure to predict protein function. For example, ContactPFP[24] predicts protein functions through contact map alignment, and DeepFRI[25] is a convoluted neural network model trained

with contact map and a protein language model. However, these methods use the full structures rather than structural sites that could reveal insights into ligand binding. Furthermore, the accuracy of their prediction is limited by poor annotation within the databases.

Currently, there are limited computational tools that can detect lipid interacting proteins precisely. González-Díaz et al. developed LIBP-Pred[26] to predict lipid binding proteins by using the electrostatic potential of residues within a coarse segmentation of the protein. However, LIBP-Pred is limited to experimental structures and does not first determine the putative ligand binding sites within the protein. Furthermore, the method does not tolerate disordered regions within proteins due to its use of coarse segmentation. A second predictor, MBPpred[27], was reported by Nastou et al. This method predicts membrane binding proteins using profile hidden Markov models. However, MBPpred often predicts membrane protein interactions with lipids driven by the hydrophobic surfaces of the protein embedded within lipid bilayers. Finally, Katuwawala et al. developed DisoLipPred[28], a multi-tool predictor where the tool first identifies disordered regions within the protein and the second tool uses neural networks to predict the probability that the disordered residues interact with lipids.

Even with these tools, the key challenge is a poor understanding of essential drivers of molecular recognition between proteins and lipids. There are numerous examples of distinct proteins that recognize the same lipid, and examples of lipid transport proteins with broad substrate scopes. It appears that, to an extent, hydrophobic interactions with amino acids are important for stabilizing the hydrophobic tails of lipids, and hydrogen bonding may be necessary for engagement with the polar heads of the lipids. Given that multiple amino acids can participate in hydrophobic interactions and hydrogen bonding, it is difficult to assign protein motifs that enable the recognition of lipids.

We have an interest in identifying lipid binding proteins in the proteomes of pathogenic bacteria that acquire lipids from their hosts. Focusing on sterol lipids, we realized that there are few proteins in bacteria with canonical sterol sensing domains, despite the fact that pathogenic bacteria acquire host sterols[9], commensal gut microbes modify host cholesterol[29-31], and primitive bacteria make sterols[32,33]. This finding suggested that bacteria may have evolved a machinery for handling sterols that is distinct from those found in eukaryotes. Reports of a divergence in sterol synthesis in the bacterial domain lent credence to the idea[32], and our recent

identification of novel sterol binding domains in bacteria further supported this hypothesis[34]. As anticipated, the molecular recognition of sterol lipids by bacterial proteins is not mediated by a particular class of amino acids but by amino acids with shared physical and chemical properties. We reasoned that these features could be used to detect the presence of lipid binding sites in protein structures.

Because there are currently no efficient structural bioinformatics tools to spur the discovery of novel lipid interacting proteins on proteome-wide scales, we developed SLiPP (Structure-based Lipid-interacting Pocket Predictor; Fig. 1). SLiPP works by identifying ligand binding pockets within experimental and computational protein structures (predicted by AlphaFold) and uses a machine learning model to detect physiochemical features consistent with lipid binding sites. By focusing on physiochemical features, SLiPP eliminates the reliance on sequence or protein folds and in doing so avoids biasing the discovery to well-characterized lipid binding domains. The approach was used to reveal the putative lipid interactome in the proteomes of *Escherichia coli* (*E. coli*), *Saccharomyces cerevisiae* (yeast), and *Homo sapiens* (human), as well as select pathogenic bacteria (Tables S1-S3). The predicted outputs in well characterized proteomes are validated by gene ontology enrichment analysis. We posit that this "easy-to-use" tool will vastly accelerate the pace at which novel lipid binding domains are discovered.

Results

Physiochemical properties of ligand pockets

A set of lipid ligands and non-lipid ligands were selected from the ligand bound protein structures in PDB data base to extract the ligand binding site information. For lipids, structures of proteins bound to cholesterol (CLR), myristic acid (MYR), palmitic acid (PLM), stearic acid (STE), and oleic acid (OLA) were selected (Supporting file 1). These lipids were chosen because there were at least 20 entries of each, which we posit is important to allow a degree of generalization of their ligand binding pockets. Phospholipids, sphingolipids, and glycerolipids were not used due to the limited number of available structures, but their structural similarity to the selected lipids should ensure that proteins recognizing them can still be identified by SLiPP. For non-lipid entries, representatives from each primary metabolite group were selected: adenosine (AND) for nucleosides, β-D-glucose (BGC) for saccharides, and cobalamin (B12) and coenzyme A (COA) for common cofactors. The ligand binding pockets were

identified using the dpocket module of fpocket,[35,36] which typically predicts pockets with ligands ("true" pockets) in addition to pockets that share no overlap with the ligand ("pseudo-pockets"). Dpocket uses 17 properties as pocket descriptors, and these descriptors can be divided into 4 categories: size-related, hydrophobicity-related, alpha sphere-related, and miscellaneous. To understand the physiochemical space of ligand binding pockets, we performed principal component analyses (PCA) on all the pockets extracted from the PDB structures (Fig. 2).

The PCA (Fig. 2A) reveals a clear separation of lipid binding pockets (LBPs) from non-lipid binding pockets (nLBPs) and pseudo-pockets (PPs), suggesting that it is possible to build a classifier that describes LBPs. The difference between LBPs and nLBPs is more pronounced in the second principal component (PC2), which is dominated by hydrophobicity-related properties (Fig. 2C). The observation is anticipated, as the key distinguishing feature between the lipid ligands and non-lipid ligands in this dataset pertains to hydrophobicity (Fig. S1); this characteristic is also reflected in the amino acid composition of ligand binding pockets. When comparing PPs and ligand binding pockets, we also observed a difference in the first principal component (PC1), with PPs exhibiting lower values compared to the ligand binding pockets (Fig. 2C). The clear distinction of size and hydrophobicity-related properties for the three classes of pockets should permit the use of machine learning to create a classifier for LBPs. However, no distinction is apparent when considering the individual lipids, which suggests that the pocket properties described by fpocket are not sufficient to detect differences between the selected lipids (Fig. 2B).

Construction of a classifier

Deep learning has gained a lot of attention due to potential applications in biomedicine. However, it requires significantly large datasets that are not available for many ligand-protein interaction. As such, we built a classifier using machine learning, where the algorithm first learns patterns from the dataset and then uses those patterns to make predictions. This approach is ideal for small datasets where generalizable patterns exist. To build the classifier, we first identified a suitable machine learning algorithm for the dataset (Fig. 3A). Six commonly used algorithms were tested: support vector machine (SVM), logistic regression (Log), k-nearest neighbors (kNN), naïve bayes (NB), decision tree (DT), random forest (RF). The performance of each algorithm was assessed in 25 independent iterations of stratified shuffle sampling. To

assess the performance, 6 metrics were calculated: area under receiver operating curve (AUROC), accuracy, F1 score, specificity, sensitivity, and precision (see Methods). These tests revealed RF performed best with a F1 score of 0.775, area under receiver operating curve (AUROC) of 0.980, and accuracy of 99.1%. Following these tests, the RF algorithm was selected to construct our classifier because of its performance. Because of the highly imbalanced nature of the dataset (vide infra), the sensitivity for all algorithms is low (ranging from 42.0% to 68.2%) while the specificity is much higher (ranging from 95.0% to 99.9%). Of note is that naïve bayes is the least-performing algorithm for our dataset because of the large number of false positives it produces. This may be due to the fact that naïve bayes is a probabilistic model, and its use with highly imbalanced datasets like ours makes it such that the prior probabilities severely affect the posterior probability.

The pocket detection algorithm has the propensity to detect a high number of PPs (relative to LBPs and nLBPs), thereby leading to a highly imbalanced dataset which could decrease the sensitivity of the classifier. The original full dataset contains 1,783 LBPs, 3,000 nLBPs, and 70,644 PPs. In an attempt to test whether the sensitivity could be improved, we sampled different numbers of PPs to include in the dataset; this produced datasets with various levels of imbalance (Fig. 3B). To correctly assess the effect of these more balanced datasets, the performance was cross validated using a set of test data not previously used in the training dataset. With this new approach, a dataset with 20-fold more PPs than LBPs performed the best, revealing an accuracy of 99.2% and F1 score of 0.797. The dataset with 5-fold more PPs and the full unbalanced dataset performed similarly to the 20-fold dataset, while the equally balanced dataset performed the worst. The precision scores followed a trend similar to the accuracy and F1 scores, wherein the full unbalanced dataset yielded a more precise model, and the equally balanced dataset had a precision score of 42.9%. As anticipated, use of the highly imbalanced dataset lowers the sensitivity (85.9% in the equally balanced dataset and 69.2% in the full dataset). Together, these results suggest that the classifier model has learnt the population distribution of LBPs and PPs from an unbalanced dataset and therefore classifies more pockets, including the majority of PPs, as nLBPs. SLiPP is much more stringent when classifying LBPs, and this is reflected in the high precision and specificity scores. On the other hand, the predictor is more forgiving when trained with a more balanced dataset, leading to high sensitivity but low precision. Because the classifier was created as a tool to spur the

discovery of novel lipid binding proteins, we reasoned that it is best to prioritize the sensitivity over the high F1 and specificity scores. This would thereby produce more hits that can be validated with additional bioinformatics and biochemical methods. Because of this, the dataset with five-fold excess of PPs was used to train and optimize the model.

A second approach to improve the model's performance focused on fine-tuning its hyperparameters (see Methods). To do this, we first performed a random search on hyperparameters to maximize the F1 score. Following that, we did a fine grid search around the hyperparameters chosen in the previous round. While the performance was not substantially improved after two rounds of optimization (Fig. S2), the optimized hyperparameters resulted in a more computationally expensive model. Therefore, we decided to retain the default hyperparameters from the sklearn[37] package.

Performance of the classifier

After generating the classifier model, the performance was assessed with a subset of the data not used in the training of the classifier model (the independent test dataset, Fig. 4A). The model performed well with this test dataset (Table 1) and revealed a AUROC of 0.970, an accuracy of 96.8%, a F1 score of 0.869, a precision of 92.6%, and a sensitivity of 81.8%. To perform an additional assessment of the model, we assembled another dataset of ligand-free (apo) structures of select ligand binding proteins (Supporting file 2) and AlphaFold-predicted models (Supporting file 3) of ligand binding proteins. Given that these structures are all free of ligands, we used fpocket to predict pockets. All pockets predicted by fpocket were then applied to the classifier to detect proteins capable of binding lipids; these proteins would have pockets with prediction scores higher than those unlikely to bind to lipids. The detection of low scoring pockets would indicate that the protein is not a lipid binding protein.

The results from this exercise revealed the AUROC of apo PDB (Fig. 4B) and AlphaFold (Fig. 4C) datasets are 0.828 and 0.851, while the F1 scores of the two test datasets are 0.726 and 0.643. These two metrics indicate a lower performance of the classifier model when testing with the external datasets (Table 1). A closer inspection of the performance reveals that the precision was less affected (89.1% and 91.8% (compared with 92.6% in the original dataset) but the sensitivity decreased (from 81.8% to 61.2% in the PDB structures and 49.5% in the AlphaFold models). This

reduction in the performance can be explained by the inaccuracy of fpocket in accurately identifying ligand-binding pockets: the incorrect identification of ligand-binding pockets results in misclassification of the protein as a non-lipid binding protein. This inaccurate identification of pockets is more pronounced for LBPs than for nLBPs and PPs. A critical pocket identification feature of fpocket is its alpha sphere clustering algorithm. However, the algorithm sometimes separates large, continuous pockets into several small pockets that are no longer predicted to be LBPs. One example is the sterol binding protein BstC[34] (Fig. S3), where a continuous pocket is predicted to be two separate pockets that each have low SLiPP scores. Despite the reduced sensitivity, the high precision of SLiPP suggests that it could become a powerful tool to aid the discovery of novel LBPs.

Detection of putative lipid binding proteins in the *Escherichia coli, Saccharomyces cerevisiae*, and *Homo sapiens* proteomes

With a classifier model in hand, we investigated its ability to predict lipid binding proteins in several well-annotated proteomes. To do this, we leveraged AlphaFold predicted structures, as they are readily available for most proteomes. Because AlphaFold models include signal peptides that could result in inaccurate pocket detection, these moieties were identified via SignalP[38] and removed from the models prior to the prediction. To reduce the computational time, proteins containing less than 100 amino acids were filtered out, as we considered these unlikely to form sufficiently large binding pockets to accommodate lipids. Additionally, low confidence AlphaFold models (pLDDT < 70) were removed (Fig. 1b).

The *Escherichia coli* proteome

E. coli proteome has 4403 proteins. Of these, 606 were removed because of their small size, and 77 were removed because of poor confidence in the AlphaFold prediction. Of the remaining 3720 proteins, 159 proteins were assessed as having a SLiPP score of 0.5 or higher, indicating that they might bind lipids (Supporting file 4). This fraction corresponds to 3.6% of the proteome (Fig. 5A and S4). An inspection of the top ten scoring proteins revealed the presence of already annotated LBPs such as the apolipoprotein N-acyltransferase Lnt, and the phospholipid transport system MlaC (Table 2). Also in this top tier are the two ubiquinol binding proteins: cytochrome bd-II ubiquinol oxidase AppB and AppC. Given the structural similarity of ubiquinol and polar

lipids, it is plausible that the predictor detects ubiquinol sites as capable of accommodating lipids. Additionally, the top hits include three potential lipid binders that are not yet to be experimentally verified: AsmA, YceI, and YhdP (Fig. 6A-C). AsmA and YhdP are inferred to be involved in lipid homeostasis through gene deletion studies[39,40], whereas YceI is thought to be an isoprenoid-binding protein because of its similarity to TT1927b from *Thermus thermophilus* HB8; an isoprenoid-bound structure exists for TT1927b[41]. Surprisingly, there are 3 proteins of unknown function in the highest scoring tier: YajR, YfjW, and YchQ (Fig. 6D-F). A crystal structure of YajR shows that it has a morphology typical of a major facilitator superfamily transporter but has an extra C-terminal domain; the function and substrate of YajR remains unknown[42]. YfjW is a completely uncharacterized protein that shares no sequence homology with any protein family. The AlphaFold predicted model shows a unique β-taco fold for the soluble domain; this fold is seen in some lipid transport systems such as the lipopolysaccharide transport (Lpt) system and the AsmA-like proteins. YchQ is annotated to belong to the unknown function protein family SirB[43]. There are limited studies on the family; however, given that SirB is within the neighborhood of KdsA[44] (an enzyme involved in lipopolysaccharide biosynthesis), and predicted to be a membrane protein, we posit that it is a putative lipid transporter. In conclusion, SLiPP has correctly identified several well-known and putative lipid binding proteins in a well-characterized bacterial proteome and additionally hints at the function of other proteins whose roles to date remain a mystery.

The *Saccharomyces cerevisiae* proteome

In *S. cerevisiae* proteome, 416 of the 6060 proteins were filtered out because of their small size, and 1536 were excluded from the prediction due to low pLDDT scores (Supporting file 5). The prediction yielded 273 hits, which corresponds to 4.5% of the proteome (Fig. 5A and S4). A gene ontology (GO) enrichment analysis of the hits showed that the top 7 biological processes enriched are all lipid-related processes. These include transport, localization, metabolism, and biosynthesis of lipids (Fig. 5B). Interestingly, the GO terms that follow the top 7 are related to cation homeostasis and ion transport, which suggests that lipids might play a role in the regulation of ion transporters. For molecular function GO terms (Fig. 5C), the two most enriched terms are lipid transporter and O-acyltransferase activity. The analysis also showed

enrichment of heme binding proteins - while we reason that the structural resemblance of heme and lipids might be sufficient to account for this misidentification, we cannot rule out a regulatory role for heme in lipid related processes. Heme does contain a porphyrin with two carboxylic acid groups, making it somewhat amphipathic. As such, heme binding pockets might share physical-chemical features with lipid binding pockets.

The *Homo sapiens* proteome

The human proteome contains 20406 proteins. A total of 7346 proteins were excluded from the prediction due to their small sizes or low pLDDT scores (Supporting file 6). The model predicts 935 hits, or 4.6%, of the proteome as putative LBPs (Fig. 5A and S4). GO enrichment analyses similar to those performed on the yeast proteome revealed that the top 7 biological process GO terms are assigned to lipid-related processes (Fig. 5D) while the molecular function GO terms enriched are related to transport processes (Fig. 5E). Using information provided in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database[45], we observed that many of the hits are protein machineries involved in the biosynthesis of unsaturated fatty acids, steroid hormone biosynthesis, glycerolipid metabolism, and fatty acid metabolism. The data provides additional confidence that SLiPP correctly identifies lipid binding proteins annotated in public databases.

Additional analysis of the SLiPP hits from the human proteome revealed several proteins linked to diseases, including neurological, metabolic, and developmental diseases (Fig. S5). One hit is GDAP2 (SLiPP score of 0.93), a ganglioside-induced protein (Fig. 7). GDAP2 is associated with spinocerebellar ataxia, an inherited neurodegenerative disorder of cerebellum[46,47]. The protein is composed of two domains: an N-terminal macro domain and a C-terminal CRAL-TRIO domain (Fig. 7A). The macro domain is thought to be an ADP-ribose or poly(ADP-ribose) binding domain; however, the macro domain of GDAP2 is different in that it only binds to poly(A) but not ADP-ribose derivatives[48]. CRAL-TRIO domains in other proteins have been characterized as lipophilic ligand binding domains with affinity for tocopherol[49] and phospholipids[50]. However, there are few biochemical studies on GDAP2.

Within the CRAL-TRIO domain of GDAP2, SLiPP identified a LBP with a size of 2034 $Å^3$ (Fig. 7A). Most amino acids making up the pocket are hydrophobic (Fig. 7B, C and Supporting file 7). An analysis of single nucleotide polymorphisms using the

NCBI's dbSNP server revealed that the four pathogenic variants of GDAP2 have mutations in the CRAL-TRIO domain. These include the Q316 nonsense mutation, the deletion of residues 316-497, and the H400R and S436V variants (Fig. 7D) – given that these all co-localize with the SLiPP-detected lipid binding site, it is plausible that lipid binding is functionally relevant for GDAP2.

Importance of pocket descriptors in SLiPP

To understand what the model learned from the training dataset, the importance of the pocket descriptors were assessed using two methods. One measures the importance of a descriptor by calculating the mean decrease in impurities of each feature (Fig. 8A); the other assesses the importance by calculating the decrease in F1 scores when permutating each feature (Fig. 8B). Of the 17 pocket descriptors provided by fpocket, the features deemed most important are the hydrophobicity-related features. In particular, the hydrophobicity score and mean local hydrophobicity density were critical. This result is further manifested by the significant difference of hydrophobicity score and mean local hydrophobicity density of LBPs compared to the binding pockets of other ligands or PPs (Fig. S1). Interestingly, the permutation importance suggests that the surface area is the third most important feature, although we observed no obvious difference in surface area between non-lipid binding pockets and lipid binding pockets. This could be because the selected nLBPs are of ligands similar in size to the set of lipid ligands used in the training dataset.

The reliance of the classifier model on hydrophobicity provides another plausible explanation for why heme binding pockets are common false positives (*vide infra* the discussion of hits in the yeast proteome). Most heme binding pockets have hydrophobicity scores similar to both lipid and non-lipid binding pockets (Fig. 8C); a similar trend was also observed for other hydrophobicity-related parameters. This overlap is even more striking when heme binding pockets are included in the PCA plot: they evenly distribute across both lipid binding and non-lipid binding pockets (Fig. 8D). Therefore, it can be challenging to distinguish heme binding pockets from lipid binding pockets using hydrophobicity-related pocket descriptors, and inclusion of heme binding proteins in the training dataset did not resolve this problem (Fig. S6). However, a post-prediction filtering of hits using one of several heme binding protein predictors (HemeBIND[51], HeMoQuest[52], or HEMEsPred[53]) should allow the removal of heme binding proteins from the list of SLiPP hits.

Discussion and conclusion

Historically, the discovery of lipid binding proteins has been low-throughput and relied on biochemical or genetic methods. Phenotypic screens and chemoproteomics using lipid probes have enabled the identification of proteins involved in lipid binding, but these approaches are limited to culturable and/or genetically tractable organisms and require specialized equipment and expertise. Several bioinformatic-based approaches have been developed to aid the discovery of novel LBPs, but all have limitations. This has slowed the pace at which these proteins are discovered, despite their involvement in a host of critical cellular functions. SLiPP, which can detect lipid binding sites within experimental and computational three-dimensional protein structures, should accelerate the discovery of these proteins. In particular, we anticipate that the use of SLiPP will facilitate the curation of lipid binding proteins in the proteomes of pathogenic bacteria known to engage with host lipids, thereby revealing new protein targets for the curtailment of bacterial infections.

SLiPP was constructed from a classifier model that uses physiochemical properties of amino acids to identify lipid interacting pockets in proteins. By focusing on the physical and chemical properties of amino acids that make up the binding cavities, we aimed to reduce bias of the classifier for identifying proteins that share homology with well characterized lipid binding domains. As such, the classifier model may expedite the discovery of novel lipid handling domains. While the model's performance metrics are considered good, there are some notable limitations. A key one is the pocket detection algorithm's use of alpha spheres, which results in the detection of spheroid-like pockets that are large and hydrophobic. Given the amphiphilic nature of the of the ligands, and to an extent the pockets, it might be possible to train a model to extract information about the orientation of ligands. Additionally, the inclusion of higher resolution pocket descriptors could allow the model to distinguish pockets that accommodate different classes of lipids. These improvements may enhance the model's performance.

A second key limitation is that the model only identifies LBPs that are embedded within monomeric proteins. If a lipid binds to the protein's surface (such as with ApoA1 and ApoE), or if the binding site is formed upon the oligomerization of one or more subunits (an example is the recently reported MCE transport system[54]), the classifier is unable to detect these as LBPs. Regardless, the high precision suggests that SLiPP

is well suited as a tool to facilitate the identification of novel lipid interacting proteins and complements existing low-throughput discovery methods. Its ability to reveal several proteins of unknown function in the well-studied *E. coli* proteome as putative lipid binders bodes well for its utility in fueling discoveries in poorly studied organisms.

Acknowledgments

This research was supported in part by NIH grant 1R35GM150910 (to L.M.K.D.) and by the Howard Hughes Medical Institute Emerging Pathogens Initiative. C.M.D. is a Sarafan Chemistry-Biology Interface Fellow. L.M.K.D. was supported by a Terman Fellowship from Stanford University and is a MAC3 Impact Philanthropies Faculty Fellow at the Sarafan ChEM-H Institute. The authors thank Prof. Grant Rostkoff for helpful discussions.

Methods

General software and packages

The fpocket[35,36] package was used either directly in the terminal or incorporated in python under biobb_vs v4.0.0[55] package. Machine learning was accomplished with the scikit-learn v1.3.1[37] package. Other python packages used in the study are: pandas[56,57], numpy[58], matplotlib[59], seaborn[60], Biopython[61]. The AlphaFold models and fpocket outputs were visualized with PyMOL[62] and ChimeraX[63].

Construction of datasets

The PDB database was retrieved on April 27th, 2023. The training dataset was composed of four different sets of pockets: 1) pseudo-pockets (PP), 2) non-lipid binding pockets (nLBPs), 3) lipid binding pockets (LBPs), and 4) heme binding pockets. PDB entries having adenosine (ADN), cobalamin (B12), β-D-glucose (BGC), coenzyme A (COA) as standalone ligands in proteins were retrieved to extract the non-lipid binding pockets. PDB entries having cholesterol (CLR), myristic acid (MYR), palmitic acid (PLM), stearic acid (STE), oleic acid (OLA) as standalone (i.e., not covalently bound) ligands were retrieved to extract lipid binding pocket. To eliminate the possibility of identifying surface-bound lipids, structures having fewer than 10 residues within 8 Å of the ligand center-of-mass were filtered out. PDB entries with hemes (HEM) as standalone ligands were retrieved to extract heme binding pocket. The dpocket module from the fpocket package was used to extract ligand pockets in these ligand-bound structures. The pockets were defined by 17 descriptors: pocket volume (pock_vol), number of alpha spheres (nb_AS), pocket surface area (surf_vdw), pocket polar surface area (surf_pol_vdw), pocket apolar surface area (surf_apol_vdw), hydrophobicity score (hydrophobicity_score), mean local hydrophobic density (mean_loc_hyd_dens), proportion of apolar alpha sphere (apol_as_prop), proportion of polar atoms (prop_polar_atm), mean alpha sphere solvent accessibility (mean_as_solv_acc), alpha sphere density (as_dens), maximum distance of alpha spheres (as_max_dst), volume score (volume_score), polarity score (polarity_score), charge score (charge_score), flexibility (flex). A total of 3,333 non-lipid binding pockets were extracted from 1,006 non-lipid bound PDB structures, 1,981 lipid binding pockets were extracted from 780 lipid bound PDB structures, and 429 heme binding pockets were extracted from 240 heme bound PDB structures. While dpocket can extract the

ligand binding pockets, it also outputs unliganded pockets (identified by fpocket) – these unliganded pockets we defined as pseudo pockets. The pseudo pockets were used to train the machine learning model to assure that the classifier distinguishes lipid binding pockets from pseudo pockets predicted by fpocket. A total of 90,232 pseudo pockets were identified from 2,026 PDB structures. The full dataset was obtained by combining non-lipid binding pockets, lipid binding pockets, and pseudo pockets. To evaluate the effect of different datasets, an independent dataset was sampled from the full dataset and included heme binding pockets using stratified sampling with a 10% fraction size.

Selection of the machine learning algorithm

Six algorithms were used to in the study: support vector machine, logistic regression, k-nearest neighbors, naïve bayes, decision tree, and random forest. The selection of the final algorithm was based on an assessment of their performance with the full dataset. The cross validation was done with the stratified shuffled sampling method in sklearn with a 90:10 ratio. 25 random stratified samples were performed to measure the average performance.

Selection of dataset

Four datasets with different ratios of lipid binding pockets and pseudo-pockets are assessed. The full dataset consists of lipid binding pockets, non-lipid binding pockets, and all pseudo pockets. The five-fold and twenty-fold datasets reduced the number of pseudo pockets by sampling the pseudo pockets with five or twenty times of the number of lipid binding pockets. The balanced dataset was assembled by sampling the pseudo pockets to match the number of lipid binding pockets. Assessment of the effect of the more balanced datasets was done using the test dataset.

Tuning of the hyperparameters

The tuning was done on the five-fold dataset and aimed to maximize the F1 score. The first round of tuning was done with the random search method in sklearn. The following hyperparameters were tuned by randomly searching within the range indicated in the parentheses: number of estimators (100, 1000), maximum features (2, 4), maximum depth (10, 100), minimum samples to split (2, 10), minimum samples in

leaf nodes (1, 4), bootstrap (True, False). The search was done for 100 iterations and cross-validated with three-fold cross validation.

The second round of tuning was done with the grid search method in sklearn by examining all possible combinations of hyperparameters with the options in parentheses: number of estimators (100, 200, 400), maximum features (2, 3, 4), maximum depth (50, 70, 90), minimum samples to split (2, 5, 10), minimum samples in leaf nodes (1, 2, 4), bootstrap (True). The options were selected to calibrate the result perform of first optimization. The search was done for 100 iterations and cross-validated with three-fold cross validation. The performance of each model was assessed through cross validation, which was done with stratified shuffled sampling method in sklearn with a 90:10 ratio. 25 random stratified sampling was done to measure the average performance.

Assessment of models

All classifier models were assessed using six metrics: area under receiver operating curve (AUROC), accuracy, F1 score, sensitivity, specificity, and precision. AUROC is defined as the area under curve of the receiver operating curve where plots the sensitivity against 1-specificity at different threshold, where AUROC of 1 is the perfect classifier and AUROC of 0.5 is the worst classifier. Accuracy is defined as the proportion of correctly labeled samples to rest of the samples; F1 score is defined as the harmonic mean of sensitivity and precision, ranging from 0 to 1; sensitivity is defined as the proportion of correct classification within the positive class; specificity is defined as the proportion of correct classification within the negative class; and precision is defined as the proportion of correct classification within the predicted positive samples. The equation for calculating each metrics are defined down below. True positive (TP) is the count of LBPs correctly predicted as LBPs. True negative (TN) is the count of nLBPs correctly predicted as nLBPs. False positive (FP) is the count of nLBPs incorrectly predicted as LBPs. False negative (FN) is the count of LBPs incorrectly predicted as nLBPs.

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F1\ score = \frac{2}{\frac{1}{\text{sens.}} + \frac{1}{prec.}}$$

Workflow for proteome prediction

The AlphaFold models were downloaded from the database. (https://alphafold.ebi.ac.uk) The fasta sequences were retrieved from UniProt. The fasta sequences were then uploaded to SignalP 6.0 web server[38] to predict the existence and cleavage sites of signal peptides, which were removed from the structure models if detected. Two filters were used to reduce the computation burden: any model containing less than 100 amino acids was removed, and models with overall pLDDT scores less than 70 were eliminated. The remaining models were fed into the fpocket algorithm, and all pockets predicted by fpocket were subjected to prediction by the classifier. The pocket with the highest prediction score was reported as the score for the entire protein.

Gene ontology analysis

The gene ontology (GO) analysis was done with the ShinyGO 0.77 web server[64]. The false discovery rate (FDR) cutoff was set at 0.05. The GO terms shown for yeast were selected with top 10 FDR and the gene numbers in the term is at least 20 but no more than 1000. The GO terms shown for human were selected with top 10 FDR and the gene numbers in the term is at least 20 but no more than 2000.

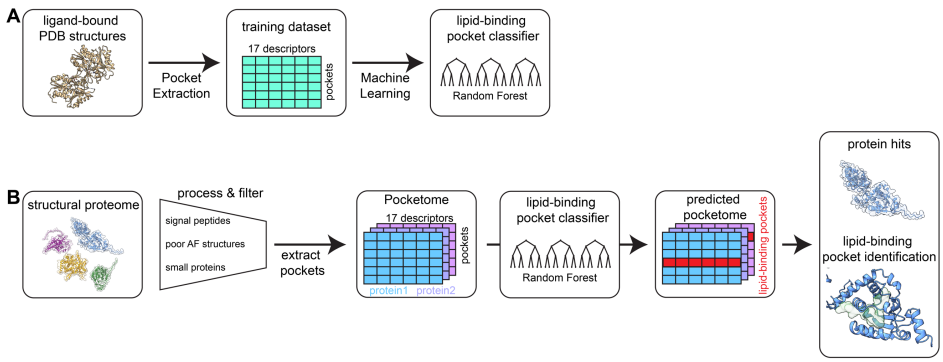

Fig. 1. (A) Description of the workflow to establish the classifier (SLiPP). (B) Workflow for proteome-wide prediction of lipid-binding pockets.

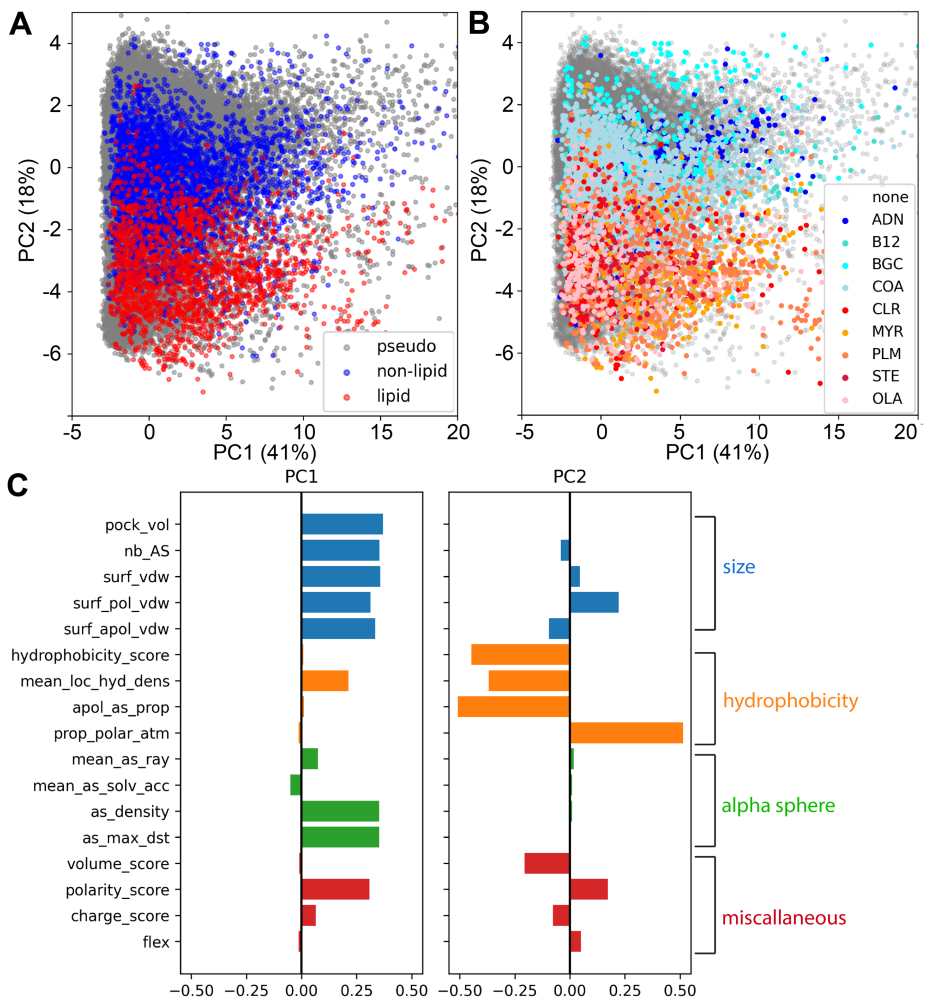

Fig. 2. PCA analyses of pockets using the 17 physiochemical properties detected by dpocket. (A, B) Score plots of the first two principal components, which describe 40.9% and 18.3% of the variance respectively. The datapoints were colored by class labels (A) and ligand identity (B). (C) A plot showing the contribution of each property to the first two principal components.

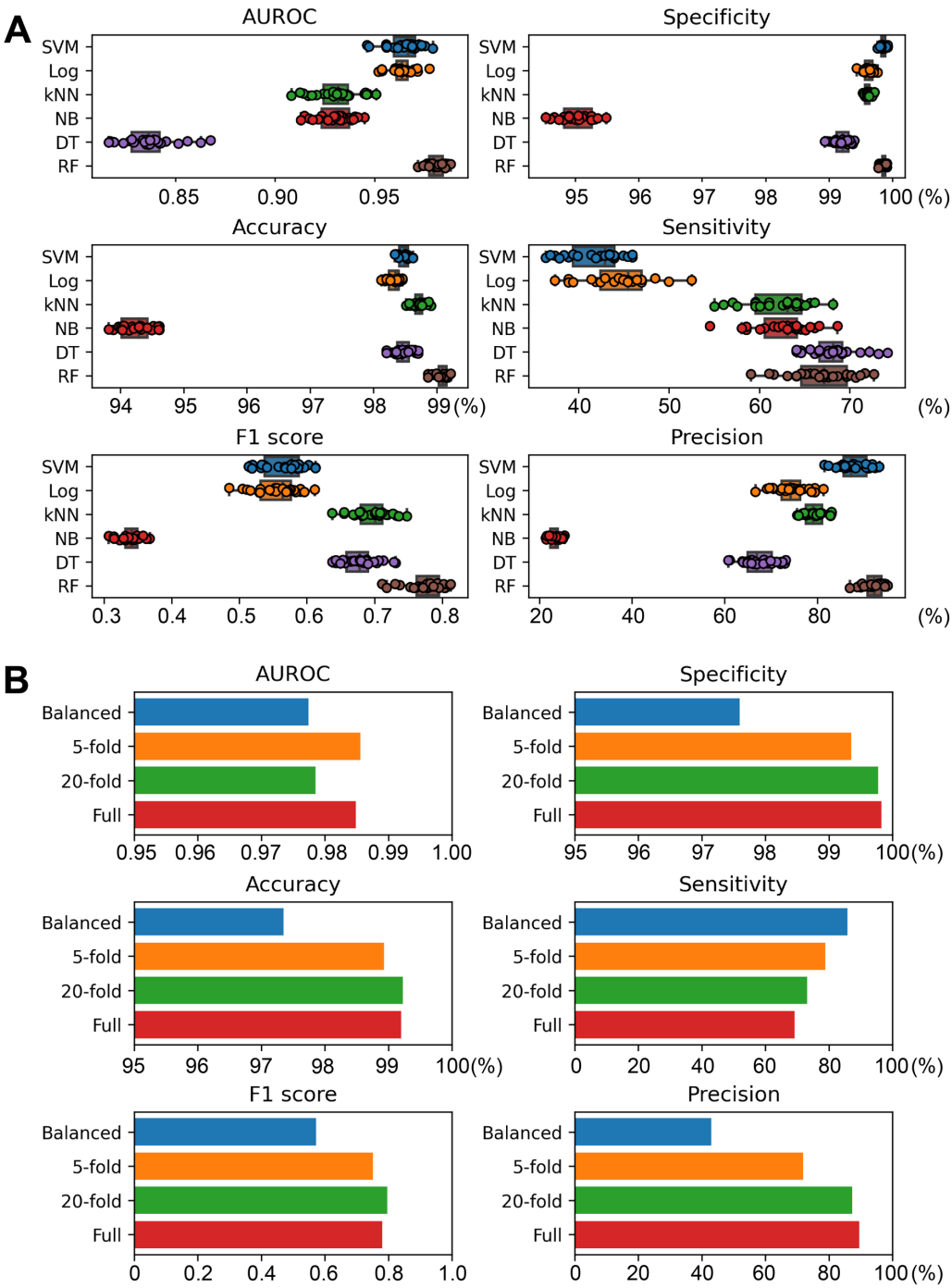

Fig. 3. (A) Assessment of machine learning algorithms used for the classifier model. The performance was assessed with 25 random seedlings. Boxes were plotted from first quartile to third quartile, while the whiskers extend to demonstrate the whole range of the data except for outliers. Outliers were defined as the datapoints outside of 1.5 times the interquartile range from the first and third quartiles. (B) Optimization of datasets for the classifier. The performance was assessed with an independent test dataset.

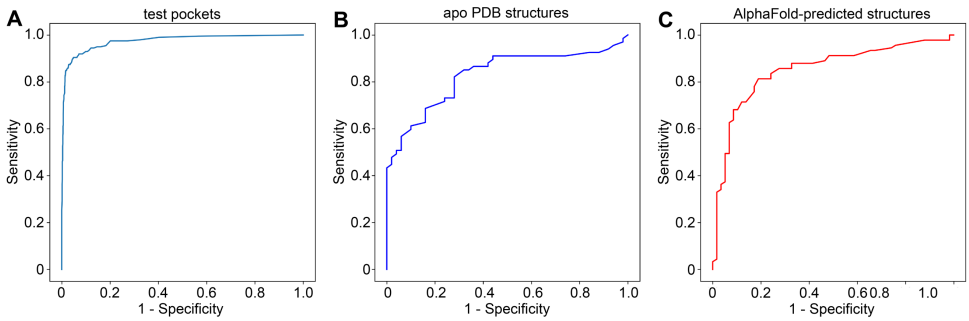

Fig. 4. Receiver operating curves for (A) the test dataset, (B) the apo PDB dataset, and (C) the AlphaFold dataset. The curves were plotted as the sensitivity vs (1 – specificity) at different thresholds.

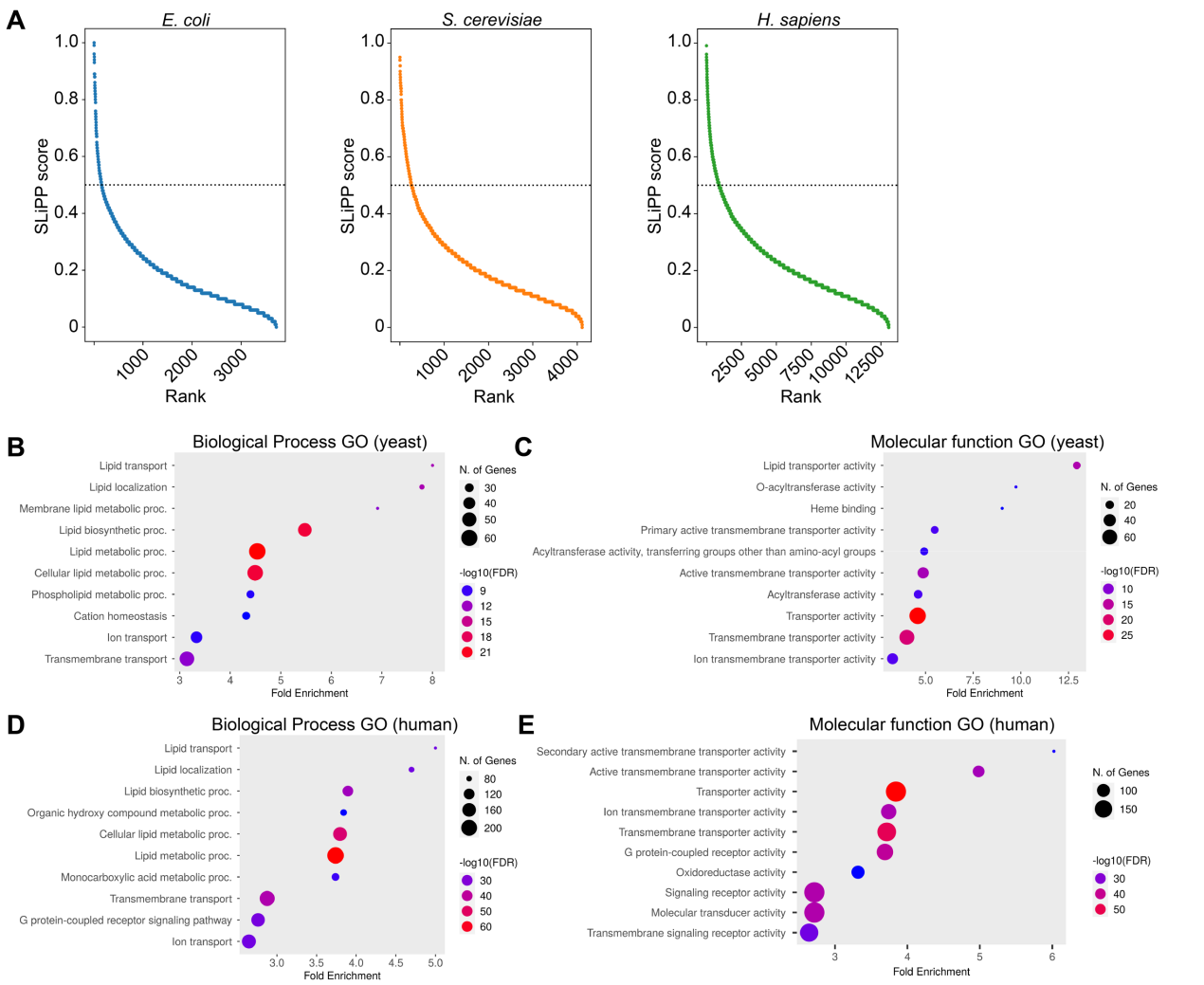

Fig. 5. (A) Prediction results of the *E. coli*, yeast, and human proteomes. The prediction scores are ranked from high to low. The dotted line indicates the prediction threshold with probability > 0.5. Gene ontology analyses of the top 10 biological process (B) and molecular function GO terms (C) in yeast and human (D, E). The size of the dot indicates the number of genes for the GO term while the color indicates the false discovery rate (FDR).

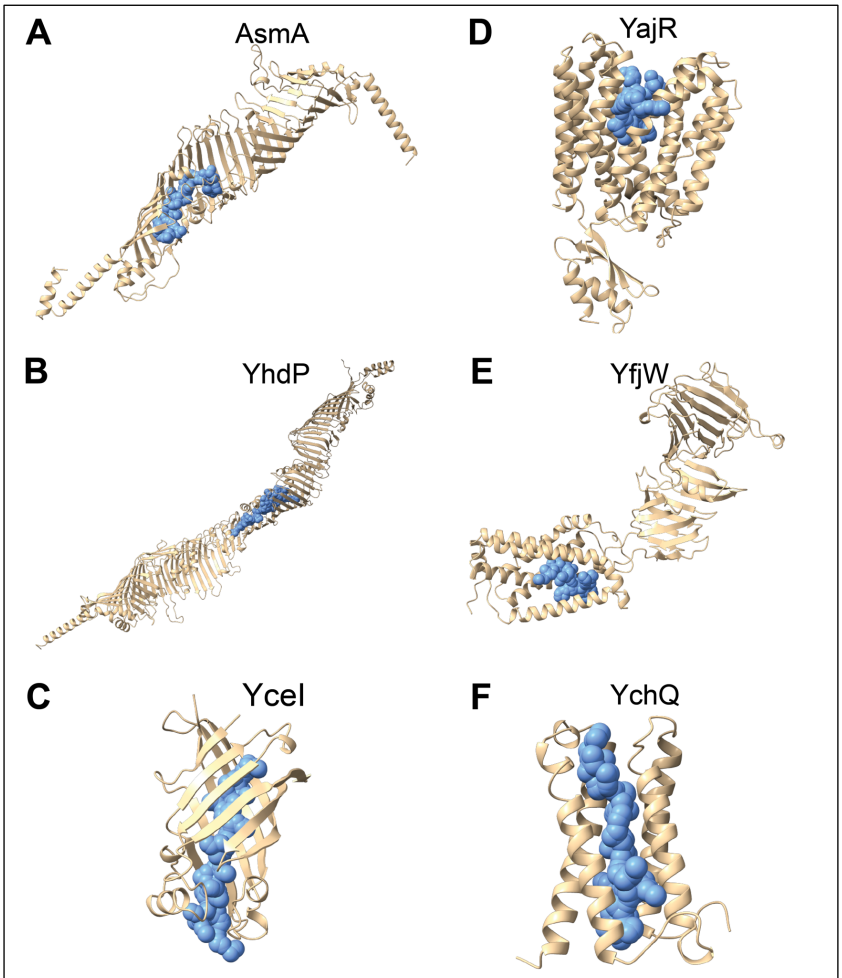

Fig. 6. SLiPP-predicted pockets (blue spheres) within the AlphaFold models of (A) AsmA (UniProt P28249), (B) YhdP (UniProt P46474), (C) YceI (UniProt P0A8X2), (D) YajR (UniProt P77726), (E) YfjW (UniProt P52138), and (F) YchQ (UniProt Q46755).

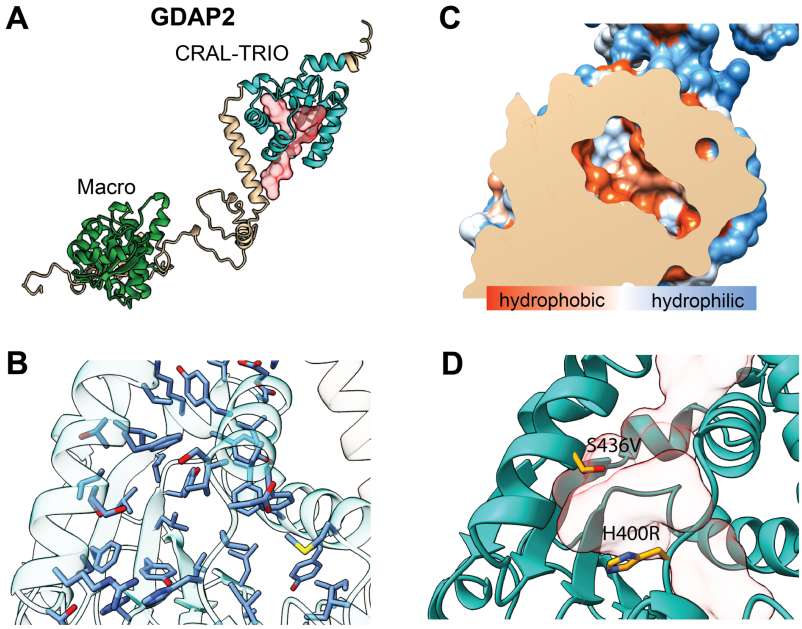

Fig. 7. SLiPP prediction analysis. (A) Domain organization of GDAP2 with the predicted lipid-binding pocket shown in red. (B) Amino acids forming the pocket are shown as blue sticks. (C) Surface depiction of the SLiPP-predicted pocket in GDAP2, displayed with hydrophobic amino acids colored in red and polar residues displayed in blue. (D) Pocket residues modified in pathogenic variants are depicted as sticks.

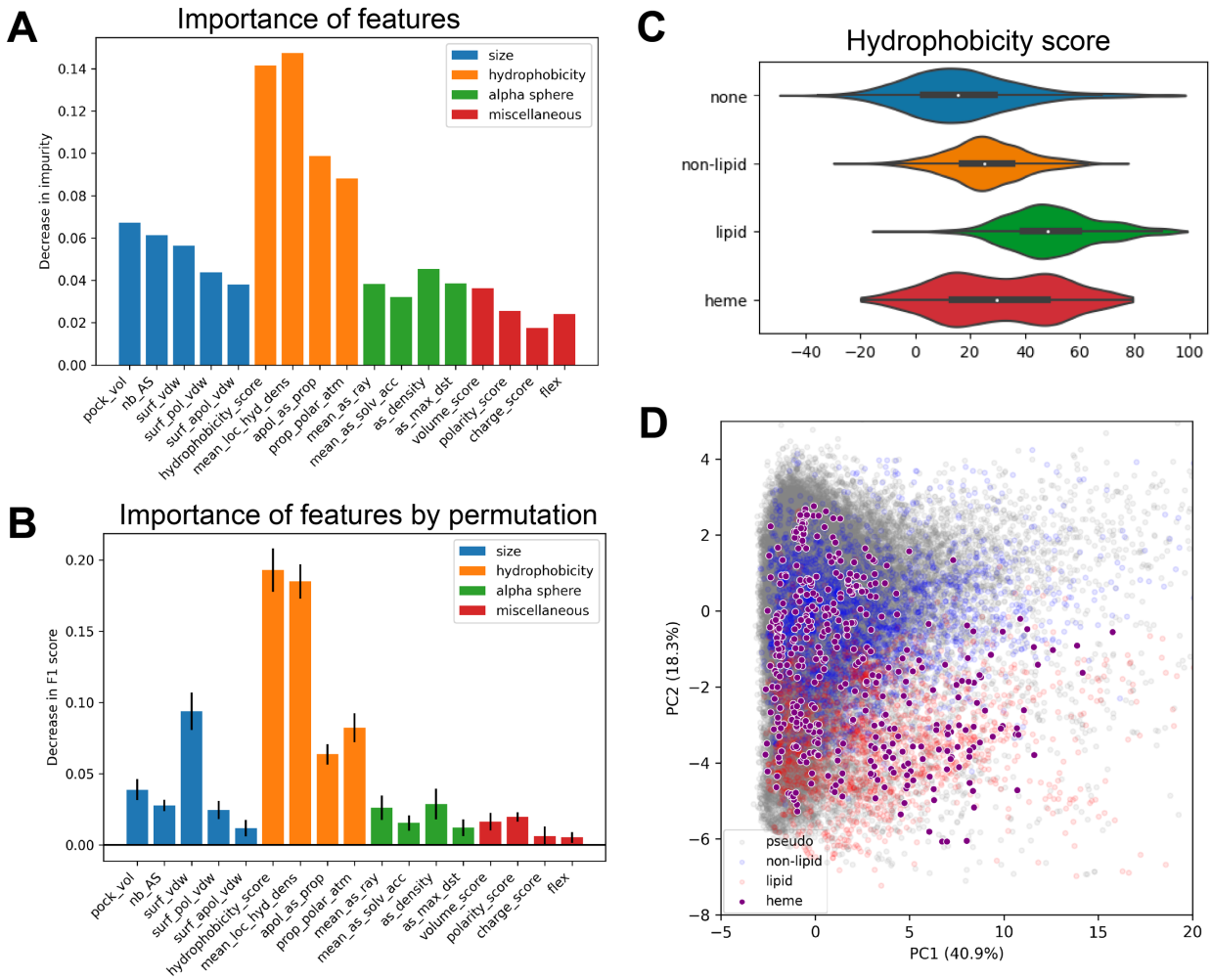

Fig. 8. The importance of pocket property was assessed by (A) the decrease in impurity and (B) the decrease in F1 score when the feature is permutated. The permutation was done in 10 repeats, with error bar indicating the standard deviation of the 10 repeats. (C) Violin plots of hydrophobicity scores with different ligand occupancies. The white dot represents median and the box plots from first quartile to third quartile. (D) Score plots on PCA analyses with the addition of heme binding pockets to the full dataset; heme binding pockets are shown as purple dots with white borders.

Table 1. Performance metrics of SLiPP from different test and validation datasets.

METRICS	Test pockets	Apo PDB	AlphaFold
AUROC	0.970	0.828	0.851
Accuracy	0.968	0.735	0.664
F1 score	0.869	0.726	0.643
Sensitivity	0.818	0.612	0.495
Precision	0.926	0.891	0.918
Cohen's kappa	0.850	0.486	0.376

Table 2. Top 10 SLiPP hits in the *E. coli* proteome. No description is provided for proteins not yet biochemically characterized.

GENE NAME	**DESCRIPTION**	**Ligand**
Lnt	Apolipoprotein N-acyltransferase	glycerophospholipid
AsmA		likely phospholipid
YceI		likely isoprenoid
AppC	Cytochrome bd-II ubiquinol oxidase	ubiquinol
MlaC	Phospholipid transport system	phospholipid
YajR		unknown
YfjW		unknown
AppB	Cytochrome bd-II ubiquinol oxidase	ubiquinol
YhdP		likely phospholipid
YchQ		unknown

References

1 Vanier, M. & Millat, G. Niemann–Pick disease type C. *Clin. Genet.* **64**, 269-281 (2003). https://doi.org:https://doi.org/10.1034/j.1399-0004.2003.00147.x

2 Yu, F. P. S., Amintas, S., Levade, T. & Medin, J. A. Acid ceramidase deficiency: Farber disease and SMA-PME. *Orphanet J. Rare Dis.* **13**, 121 (2018). https://doi.org:10.1186/s13023-018-0845-z

3 Jefferies, J. L. Barth syndrome. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics* **163**, 198-205 (2013). https://doi.org:https://doi.org/10.1002/ajmg.c.31372

4 Marshall, W. C., Ockenden, B. G., Fosbrooke, A. S. & Cumings, J. N. Wolman's disease. A rare lipidosis with adrenal calcification. *Arch. Dis. Child.* **44**, 331-341 (1969). https://doi.org:10.1136/adc.44.235.331

5 Santos, C. R. & Schulze, A. Lipid metabolism in cancer. *The FEBS Journal* **279**, 2610-2623 (2012). https://doi.org:https://doi.org/10.1111/j.1742-4658.2012.08644.x

6 Heaton, N. S. & Randall, G. Multifaceted roles for lipids in viral infection. *Trends Microbiol.* **19**, 368-375 (2011). https://doi.org:10.1016/j.tim.2011.03.007

7 Adams, F. G. *et al.* To Make or Take: Bacterial Lipid Homeostasis during Infection. *mBio* **12**, 10.1128/mbio.00928-00921 (2021). https://doi.org:doi:10.1128/mbio.00928-21

8 van der Meer-Janssen, Y. P. M., van Galen, J., Batenburg, J. J. & Helms, J. B. Lipids in host–pathogen interactions: Pathogens exploit the complexity of the host cell lipidome. *Prog. Lipid Res.* **49**, 1-26 (2010). https://doi.org:https://doi.org/10.1016/j.plipres.2009.07.003

9 Crowley, J. T. *et al.* Lipid Exchange between Borrelia burgdorferi and Host Cells. *PLOS Pathogens* **9**, e1003109 (2013). https://doi.org:10.1371/journal.ppat.1003109

10 Johnson, R. C., Livermore, B. P., Jenkin, H. M. & Eggebraten, L. Lipids of *Treponema pallidum* Kazan 5. *Infection and Immunity* **2**, 606-609 (1970). https://doi.org:doi:10.1128/iai.2.5.606-609.1970

11 Cocchiaro, J. L., Kumar, Y., Fischer, E. R., Hackstadt, T. & Valdivia, R. H. Cytoplasmic lipid droplets are translocated into the lumen of the *Chlamydia trachomatis* parasitophorous vacuole. *Proceedings of the National Academy of Sciences* **105**, 9379-9384 (2008). https://doi.org:doi:10.1073/pnas.0712241105

12 Elwell, C. A. & Engel, J. N. Lipid acquisition by intracellular Chlamydiae. *Cell. Microbiol.* **14**, 1010-1018 (2012). https://doi.org:https://doi.org/10.1111/j.1462-5822.2012.01794.x

13 Niphakis, Micah J. *et al.* A Global Map of Lipid-Binding Proteins and Their Ligandability in Cells. *Cell* **161**, 1668-1680 (2015). https://doi.org:https://doi.org/10.1016/j.cell.2015.05.045

14 Cheng, Y.-S. *et al.* A proteome-wide map of 20(S)-hydroxycholesterol interactors in cell membranes. *Nat. Chem. Biol.* **17**, 1271-1280 (2021). https://doi.org:10.1038/s41589-021-00907-2

15 Pei, Y. *et al.* Integrated lipidomics and RNA sequencing analysis reveal novel changes during 3T3-L1 cell adipogenesis. *PeerJ* **10**, e13417 (2022). https://doi.org:10.7717/peerj.13417

16 Gao, X. *et al.* RNA-Seq and UHPLC-Q-TOF/MS Based Lipidomics Study in Lysiphlebia japonica. *Sci. Rep.* **8**, 7802 (2018). https://doi.org:10.1038/s41598-018-26139-4

17 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Bio.* **215**, 403-410 (1990). https://doi.org:https://doi.org/10.1016/S0022-2836(05)80360-2
18 Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755-763 (1998). https://doi.org:10.1093/bioinformatics/14.9.755
19 Sanderson, T., Bileschi, M. L., Belanger, D. & Colwell, L. J. ProteInfer, deep neural networks for protein functional inference. *eLife* **12**, e80942 (2023). https://doi.org:10.7554/eLife.80942
20 Cao, R. *et al.* ProLanGO: Protein Function Prediction Using Neural Machine Translation Based on a Recurrent Neural Network. *Molecules* **22**, 1732 (2017).
21 Kulmanov, M., Khan, M. A. & Hoehndorf, R. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics* **34**, 660-668 (2017). https://doi.org:10.1093/bioinformatics/btx624
22 Kulmanov, M. & Hoehndorf, R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics* **36**, 422-429 (2019). https://doi.org:10.1093/bioinformatics/btz595
23 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021). https://doi.org:10.1038/s41586-021-03819-2
24 Kagaya, Y., Flannery, S. T., Jain, A. & Kihara, D. ContactPFP: Protein Function Prediction Using Predicted Contact Information. *Frontiers in Bioinformatics* **2** (2022). https://doi.org:10.3389/fbinf.2022.896295
25 Gligorijević, V. *et al.* Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.* **12**, 3168 (2021). https://doi.org:10.1038/s41467-021-23303-9
26 González-Díaz, H. *et al.* LIBP-Pred: web server for lipid binding proteins using structural network parameters; PDB mining of human cancer biomarkers and drug targets in parasites and bacteria. *Mol. Biosyst.* **8**, 851-862 (2012). https://doi.org:10.1039/C2MB05432A
27 Nastou, K. C., Tsaousis, G. N., Papandreou, N. C. & Hamodrakas, S. J. MBPpred: Proteome-wide detection of membrane lipid-binding proteins using profile Hidden Markov Models. *Biochim. Biophys. Acta* **1864**, 747-754 (2016). https://doi.org:https://doi.org/10.1016/j.bbapap.2016.03.015
28 Katuwawala, A., Zhao, B. & Kurgan, L. DisoLipPred: accurate prediction of disordered lipid-binding residues in protein sequences with deep recurrent networks and transfer learning. *Bioinformatics* **38**, 115-124 (2021). https://doi.org:10.1093/bioinformatics/btab640
29 Yao, L. *et al.* A biosynthetic pathway for the selective sulfonation of steroidal metabolites by human gut bacteria. *Nature Microbiology* **7**, 1404-1418 (2022). https://doi.org:10.1038/s41564-022-01176-y
30 Le, H. H., Lee, M.-T., Besler, K. R., Comrie, J. M. C. & Johnson, E. L. Characterization of interactions of dietary cholesterol with the murine and human gut microbiome. *Nature Microbiology* **7**, 1390-1403 (2022). https://doi.org:10.1038/s41564-022-01195-9
31 Kenny, D. J. *et al.* Cholesterol Metabolism by Uncultured Human Gut Bacteria Influences Host Cholesterol Level. *Cell Host Microbe* **28**, 245-257.e246 (2020). https://doi.org:https://doi.org/10.1016/j.chom.2020.05.013

32 Lee, A. K. *et al.* C-4 sterol demethylation enzymes distinguish bacterial and eukaryotic sterol synthesis. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 5884-5889 (2018). https://doi.org:10.1073/pnas.1802930115
33 Lee, A. K., Wei, J. H. & Welander, P. V. De novo cholesterol biosynthesis in bacteria. *Nat. Commun.* **14**, 2904 (2023). https://doi.org:10.1038/s41467-023-38638-8
34 Zhai, L. *et al.* (eLife Sciences Publications, Ltd, 2023).
35 Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics* **10**, 168 (2009). https://doi.org:10.1186/1471-2105-10-168
36 Schmidtke, P., Guilloux, V. L., Maupetit, J. & Tufféry, P. fpocket: online tools for protein ensemble pocket detection and tracking. *Nucleic Acids Res.* **38**, W582-589 (2010). https://doi.org:10.1093/nar/gkq383
37 Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825-2830 (2011).
38 Teufel, F. *et al.* SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.* **40**, 1023-1025 (2022). https://doi.org:10.1038/s41587-021-01156-3
39 Douglass, M. V., McLean, A. B. & Trent, M. S. Absence of YhdP, TamB, and YdbH leads to defects in glycerophospholipid transport and cell morphology in Gram-negative bacteria. *PLoS Genet.* **18**, e1010096 (2022). https://doi.org:10.1371/journal.pgen.1010096
40 Ruiz, N., Davis, R. M. & Kumar, S. YhdP, TamB, and YdbH Are Redundant but Essential for Growth and Lipid Homeostasis of the Gram-Negative Outer Membrane. *mBio* **12**, e02714-02721 (2021). https://doi.org:doi:10.1128/mBio.02714-21
41 Handa, N. *et al.* Crystal structure of a novel polyisoprenoid-binding protein from *Thermus thermophilus* HB8. *Protein Sci.* **14**, 1004-1010 (2005). https://doi.org:https://doi.org/10.1110/ps.041183305
42 Jiang, D. *et al.* Structure of the YajR transporter suggests a transport mechanism based on the conserved motif A. *Proceedings of the National Academy of Sciences* **110**, 14664-14669 (2013). https://doi.org:doi:10.1073/pnas.1308127110
43 Rakeman, J. L., Bonifield, H. R. & Miller, S. I. A HilA-Independent Pathway to *Salmonella typhimurium* Invasion Gene Transcription. *J. Bacteriol.* **181**, 3096-3104 (1999). https://doi.org:doi:10.1128/jb.181.10.3096-3104.1999
44 Strohmaier, H., Remler, P., Renner, W. & Högenauer, G. Expression of genes kdsA and kdsB involved in 3-deoxy-D-manno-octulosonic acid metabolism and biosynthesis of enterobacterial lipopolysaccharide is growth phase regulated primarily at the transcriptional level in *Escherichia coli* K-12. *J. Bacteriol.* **177**, 4488-4500 (1995). https://doi.org:doi:10.1128/jb.177.15.4488-4500.1995
45 Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.* **49**, D545-D551 (2020). https://doi.org:10.1093/nar/gkaa970
46 Dong, H.-L., Cheng, H.-L., Bai, G., Shen, Y. & Wu, Z.-Y. Novel GDAP2 pathogenic variants cause autosomal recessive spinocerebellar ataxia-27 (SCAR27) in a Chinese family. *Brain* **143**, e50-e50 (2020). https://doi.org:10.1093/brain/awaa121
47 Breza, M. *et al.* A homozygous GDAP2 loss-of-function variant in a patient with adult-onset cerebellar ataxia. *Brain* **143**, e49-e49 (2020). https://doi.org:10.1093/brain/awaa120

48 Neuvonen, M. & Ahola, T. Differential Activities of Cellular and Viral Macro Domain Proteins in Binding of ADP-Ribose Metabolites. *J. Mol. Bio.* **385**, 212-225 (2009). https://doi.org:https://doi.org/10.1016/j.jmb.2008.10.045
49 Panagabko, C. *et al.* Ligand Specificity in the CRAL-TRIO Protein Family. *Biochemistry* **42**, 6467-6474 (2003). https://doi.org:10.1021/bi034086v
50 Sha, B., Phillips, S. E., Bankaitis, V. A. & Luo, M. Crystal structure of the *Saccharomyces cerevisiae* phosphatidylinositol-transfer protein. *Nature* **391**, 506-510 (1998). https://doi.org:10.1038/35179
51 Liu, R. & Hu, J. HemeBIND: a novel method for heme binding residue prediction by combining structural and sequence information. *BMC Bioinformatics* **12**, 207 (2011). https://doi.org:10.1186/1471-2105-12-207
52 Paul George, A. A. *et al.* HeMoQuest: a webserver for qualitative prediction of transient heme binding to protein motifs. *BMC Bioinformatics* **21**, 124 (2020). https://doi.org:10.1186/s12859-020-3420-2
53 Zhang, J., Chai, H., Gao, B., Yang, G. & Ma, Z. HEMEsPred: Structure-Based Ligand-Specific Heme Binding Residues Prediction by Using Fast-Adaptive Ensemble Learning Scheme. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **15**, 147-156 (2018). https://doi.org:10.1109/TCBB.2016.2615010
54 Chen, J. *et al.* Structure of an endogenous mycobacterial MCE lipid transporter. *Nature* (2023). https://doi.org:10.1038/s41586-023-06366-0
55 Andrio, P. *et al.* BioExcel Building Blocks, a software library for interoperable biomolecular simulation workflows. *Scientific Data* **6**, 169 (2019). https://doi.org:10.1038/s41597-019-0177-4
56 pandas-dev/pandas: Pandas v. latest (Zenodo, 2020).
57 McKinney, W. in *Proceedings of the 9th Python in Science Conference.* (eds Stéfan van der Walt & Jarrod Millman) 56-61.
58 Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357-362 (2020). https://doi.org:10.1038/s41586-020-2649-2
59 Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* **9**, 90-95 (2007). https://doi.org:10.1109/MCSE.2007.55
60 Waskom, M. L. seaborn: statistical data visualization. *Journal of Open Source Software* **6**, 3021 (2021). https://doi.org:10.21105/joss.03021
61 Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422-1423 (2009). https://doi.org:10.1093/bioinformatics/btp163
62 Schrodinger, L. *The PyMOL molecular graphics system, version 1.8* (The PyMOL molecular graphics system, version 1.8, 2015).
63 Meng, E. C. *et al.* UCSF ChimeraX: Tools for structure building and analysis. *Protein Sci.* **32**, e4792 (2023). https://doi.org:https://doi.org/10.1002/pro.4792
64 Ge, S. X., Jung, D. & Yao, R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics* **36**, 2628-2629 (2019). https://doi.org:10.1093/bioinformatics/btz931

Supporting Information for

Rapid proteome-wide prediction of lipid-interacting proteins through ligand-guided structural genomics

Jonathan Chiu-Chun Chou,[1] Cassandra M. Decosto,[1#] Poulami Chatterjee,[1#] Laura M. K. Dassama[1,2*]

[1]Department of Chemistry and Sarafan ChEM-H Institute, Stanford University, Stanford, CA 94305

[2]Department of Microbiology and Immunology, Stanford School of Medicine, Stanford, CA 94305

[#]Equal contributors

[*]Correspondence to: dassama@stanford.edu

Supporting figures

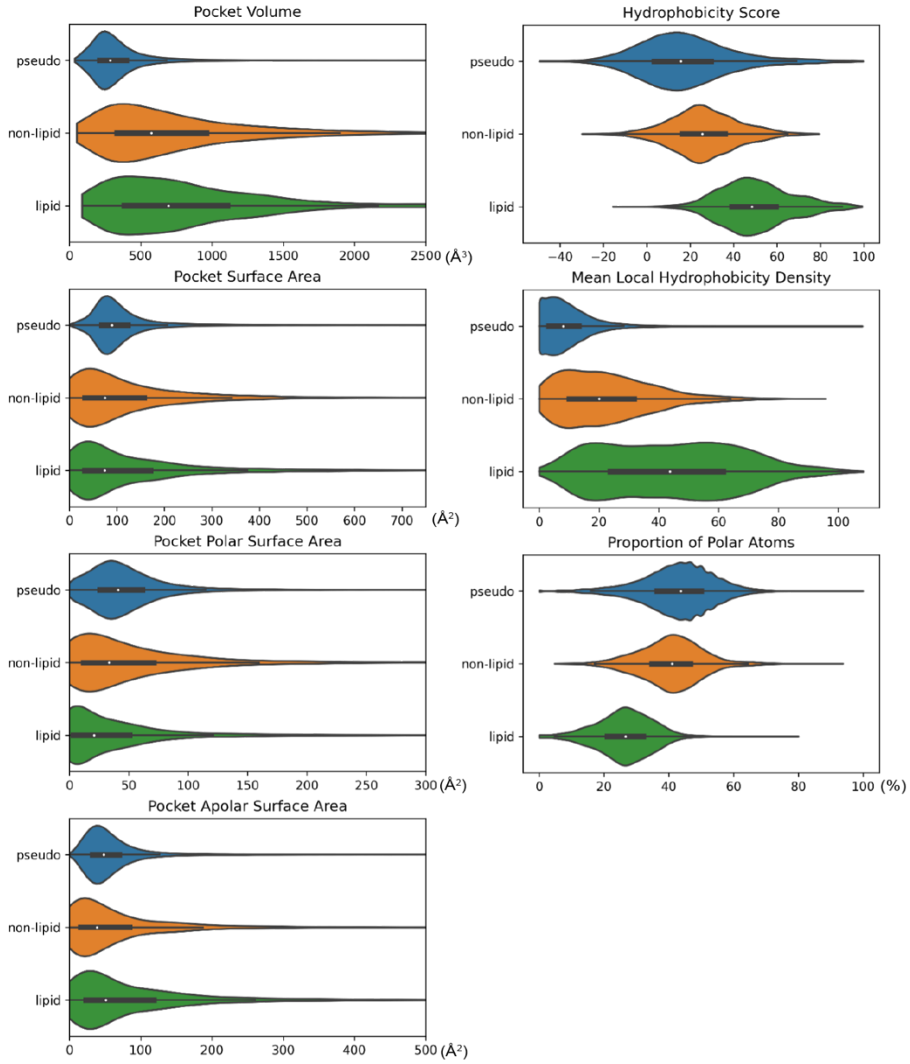

Fig. S1. Violin plots showing representative pocket descriptors of LBPs, nLBPs, and PPs. Hydrophobicity score is based on an arbitrary scale suggested by Monera et al.[1], where glycine is defined as 0 and phenylalanine as 100.

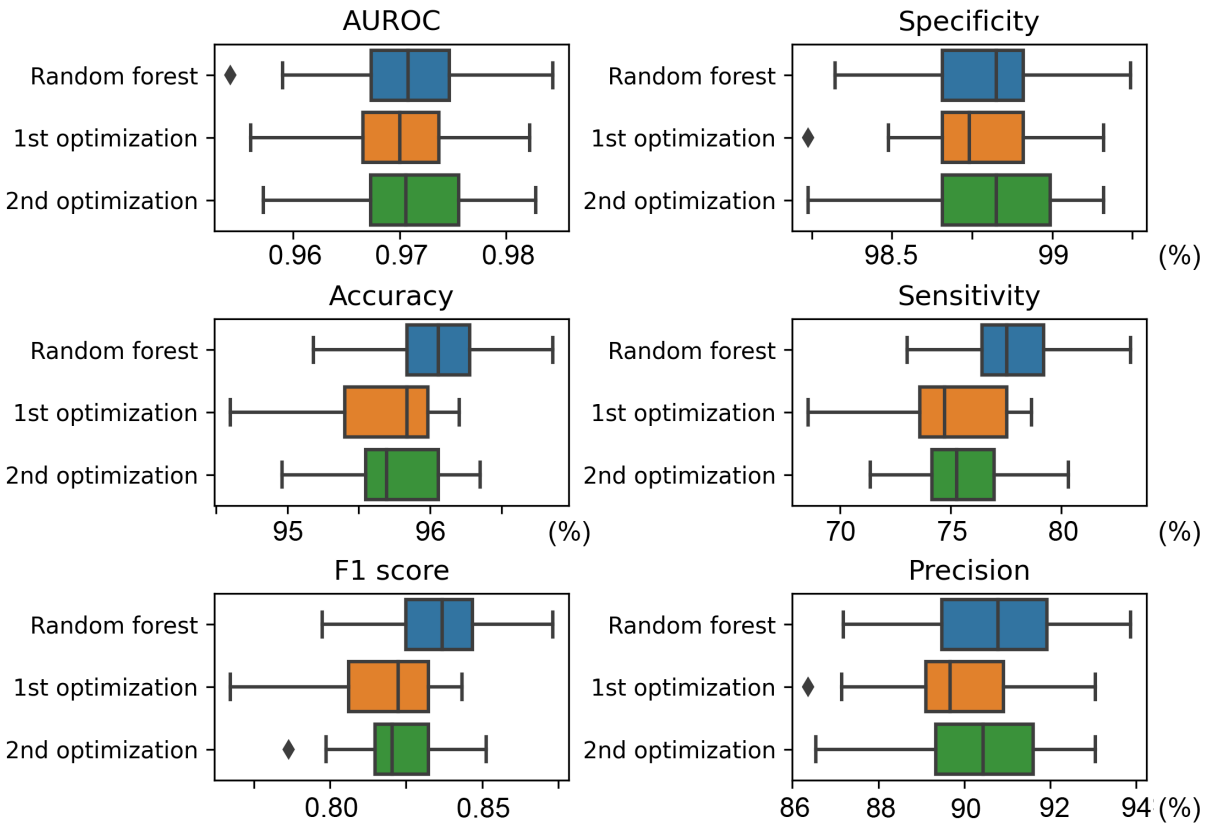

Fig. S2. Hyperparameters optimization for the classifier model. The performance was assessed with 25 random seedlings. The boxes were plotted from first quartile to third quartile, while the whiskers extend to demonstrate the whole range of the data with the exception of outliers.

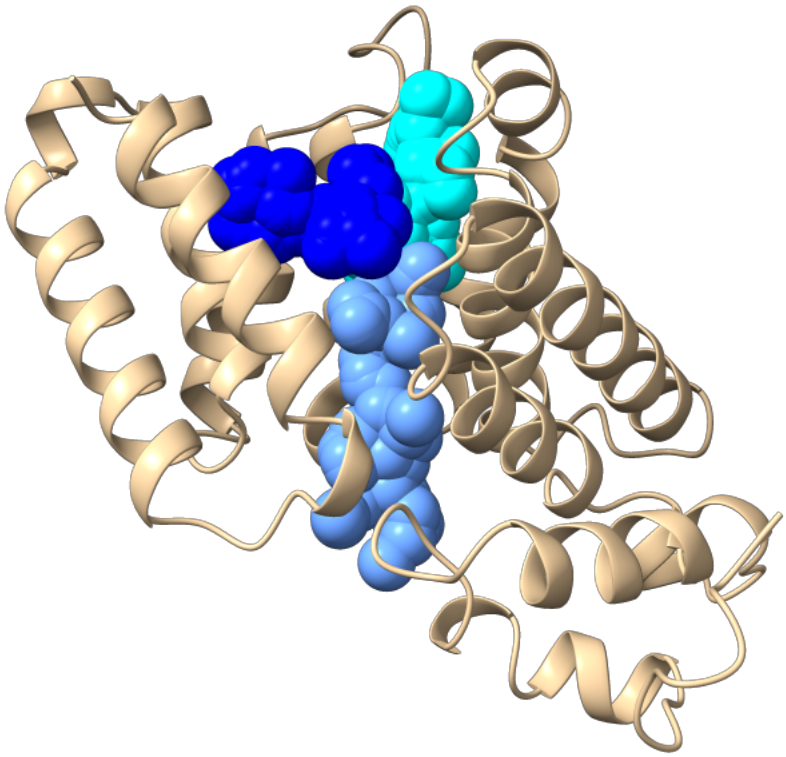

Fig. S3. Prediction of pockets within BstC (PDB 7T1S) using fpocket. AlphaFold model is colored in wheat and the predicted pockets is colored in various shades of blue to indicate separate the pockets predicted by fpocket.

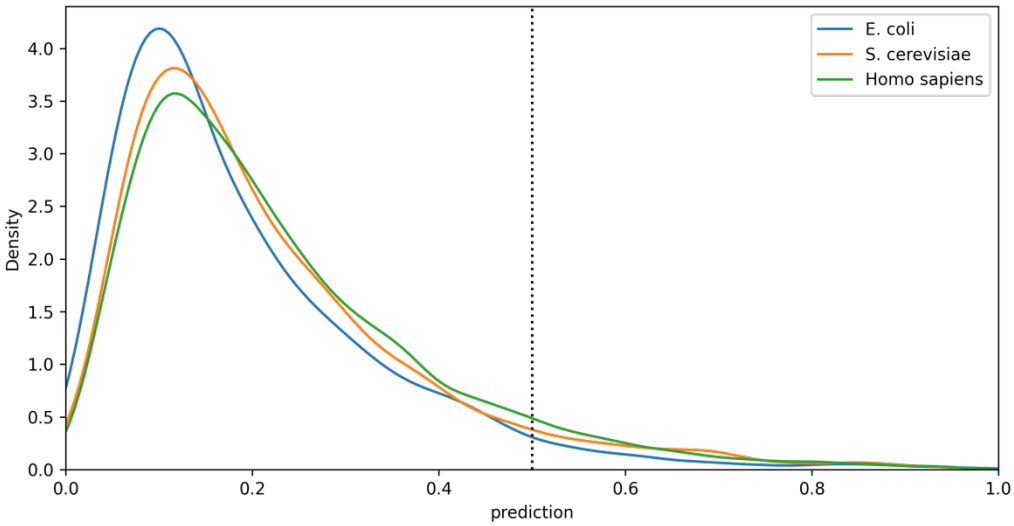

Fig. S4. The probability distribution function of SLiPP scores from the *E. coli*, yeast, and human proteomes. The dashed line indicates the hit threshold prediction score for SLiPP (0.5).

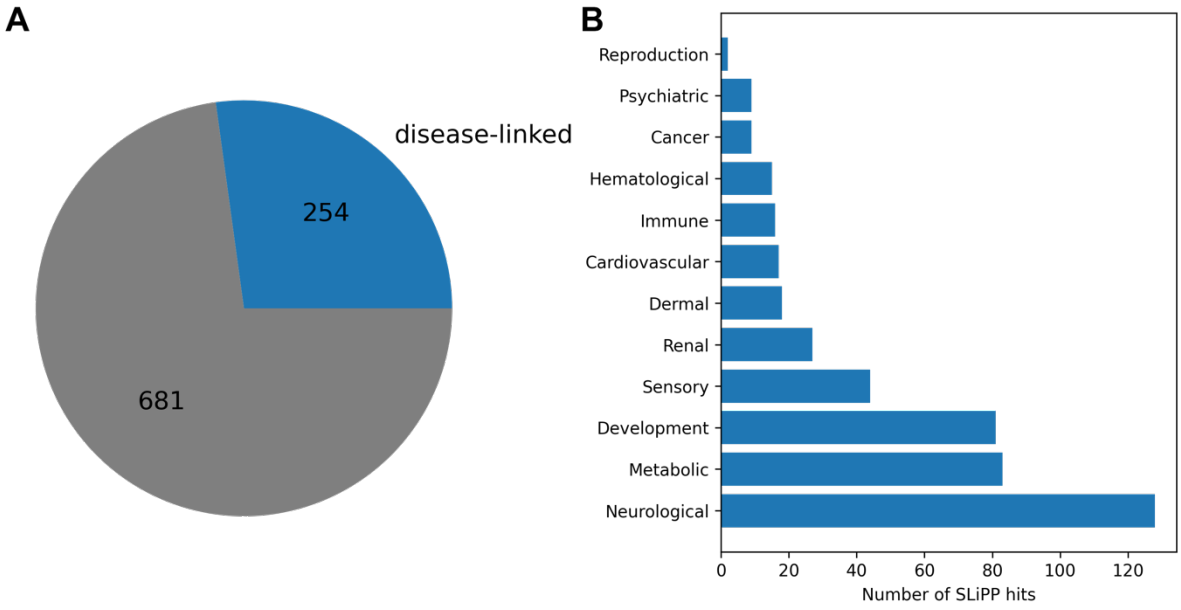

Fig. S5. SLiPP hits in the human proteome. (A) Pie chart showing the proportion of SLiPP hits that are associated with one or more diseases, as noted in UniProt. (B) Graph revealing the specific disease categories that SLiPP hits are associated with.

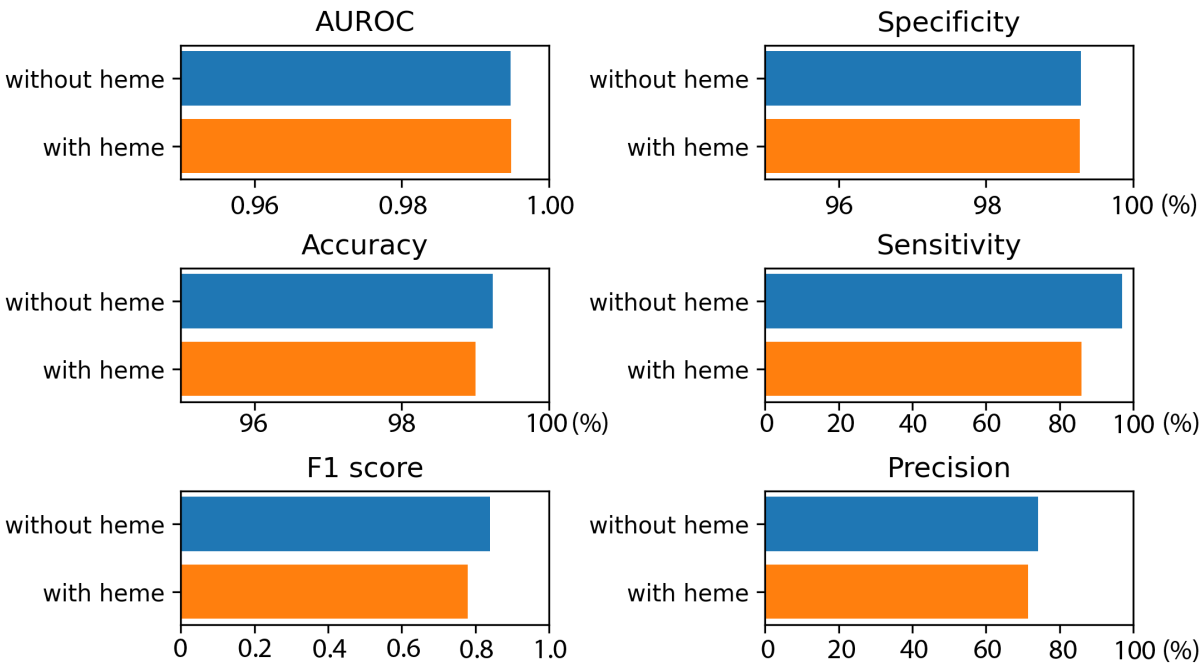

Fig. S6. Optimization of datasets with or without heme binding proteins. The performance of the classifier was assessed with a 10% independent test dataset.

Supporting tables

Table S1. Top 10 SLiPP hits of putative lipid binding proteins in *Borrelia burgdorferi* B31. Descriptions are annotations provided by UniProt.

GENE	DESCRIPTION
Lgt, BB_0362	Phosphatidylglycerol-prolipoprotein diacylglyceryl transferase
Lnt, BB_0237	Apolipoprotein N-acyltransferase
YajC, BB_0651	Sec translocon accessory complex subunit
BB_0584	Conserved hypothetical integral membrane protein
BB_0597	Methyl-accepting chemotaxis protein
BB_0368	Glycerol-3-phosphate dehydrogenase
BB_0117	UPF0073 membrane protein
BB_0747	Oligopeptide ABC transporter, permease protein
DnaJ, BB_0517	Chaperone protein
ResT, BB_B03	Telomere resolvase

Table S2. Top 10 SLiPP hits of putative lipid binding proteins in *Treponema pallidum pallidum*. Descriptions are annotations provided by UniProt.

GENE	DESCRIPTION
TP_0671	Sn-1,2-diacylglycerol cholinephosphotransferase
TP_0175	Uncharacterized protein
TP_0229	Type-4 uracil-DNA glycosylase
TP_0324	Uncharacterized protein
TP_0789	Uncharacterized protein
YidC, TP_0949	Membrane protein insertase
TP_0515	LptD C-terminal domain-containing protein
TP_0022	Uncharacterized protein
TP_0447	Uncharacterized protein
TP_0481	Uncharacterized protein

Table S3. Top 10 SLiPP hits of putative lipid binding proteins in *Chlamydia trachomatis*. Descriptions are annotations provided by UniProt.

GENE	DESCRIPTION
CT_850	Integral membrane protein
CydA, CT_013	Cytochrome Oxidase Subunit I
MenG, CT_428	Demethylmenaquinone methyltransferase
Lnt CT_534	Apolipoprotein N-
CT_131	Possible Transmembrane Protein
CT_573	Uncharacterized protein
MraY, CT_757	Phospho-N-acetylmuramoyl-pentapeptide-transferase
UppS, CT_450	Isoprenyl transferase
Aas, CT_776	Acylglycerophosphoethanolamine Acyltransferase
BrnQ, CT_554	Amino Acid (Branched) Transport

Supporting files

File 1: Table of the PDB structures used to create the full dataset
File 2: Table of the apo PDB structures used to create the validation dataset
File 3: Table of the AlphaFold structures used to create the validation dataset
File 4: SLiPP hits in the *E. coli* proteome
File 5: SLiPP hits in the *S. cerevisiae* proteome
File 6: SLiPP hits in the *H. sapiens* proteome
File 7: Table of amino acids making up the putative lipid binding pocket of GDAP2

References

1 Monera, O. D., Sereda, T. J., Zhou, N. E., Kay, C. M. & Hodges, R. S. Relationship of sidechain hydrophobicity and α-helical propensity on the stability of the single-stranded amphipathic α-helix. *J. Pept. Sci.* **1**, 319-329 (1995). https://doi.org:https://doi.org/10.1002/psc.310010507