Machine Learning Models to Interrogate Proteome-wide Cysteine Ligandabilities

Ruibin Liu, Joseph Clayton, Mingzhe Shen, and Jana Shen*
Department of Pharmaceutical Sciences, University of Maryland School of Pharmacy, Baltimore, MD 21201, United States
Received September 11, 2023; E-mail: jana.shen@rx.umaryland.edu

Abstract: Machine learning (ML) identification of covalently ligandable sites may significantly accelerate targeted covalent inhibitor discoveries and expand the druggable proteome space. Here we report the development of the tree-based models and convolutional neural networks trained on a newly curated database (LigCys3D) of over 1,000 liganded cysteines in nearly 800 proteins represented by over 10,000 X-ray structures as reported in the protein data bank (PDB). The unseen tests yielded 94% AUC (area under the receiver operating characteristic curve), demonstrating the highly predictive power of the models. Interestingly, application to the proteins evaluated by the activity-based protein profiling (ABPP) experiments in cell lines gave a lower AUC of 72%. Analysis revealed significant discrepancies in the structural environment of the ligandable cysteines captured by X-ray crystallography and those determined by ABPP. This surprising finding warrants further investigations and may have implications for future drug discoveries. We discuss ways to improve the models and project future directions. Our work represents a first step towards the ML-led integration of big genome data, structure models, and chemoproteomic experiments to annotate the human proteome space for the next-generation drug discoveries.

INTRODUCTION

Over the past two decades, targeted covalent inhibitor (TCI) discovery has become mainstream in the efforts to overcome the limitations of traditional reversible inhibitors and expand the druggable proteome space.[1–3] In the TCI design, an electrophilic functional group (also known as the warhead) is incorporated into a reversible ligand to enhance potency, selectivity, and target residence time or to inhibit a previously deemed undruggable protein, e.g., KRAS-G12C that lacks of a traditional ligandable pocket for reversible binding.[4] An irreversible and sometimes also reversible covalent bond is formed between the warhead and a nucleophilic or reactive amino acid residue in the target protein. Due to the high intrinsic nucleophilicity, cysteine has been the most popular site of covalent ligation. In the recent decade, a chemical proteomic technique called the activity-based protein profiling (ABPP)[5,6] has emerged as a linchpin technology in the rational design of TCIs on a large scale,[7–9] as ABPP can be performed with lysates or intact cells to assess ligandabilities of amino acid sites. The cysteine-directed ABPP experiments have quantified thousands of cysteines in various cell lines.[10,11]

In silico approaches are significantly faster and may complement the ABPP experiments to greatly accelerate the proteome-wide TCI discovery efforts. In recent years, covalent docking[10] and molecular dynamics (MD) based approaches[12–14] have been developed to assess cysteine reactivities and ligandabilities; however, these computationally intensive approaches cannot be scaled up to the proteome level. Recently, machine learning (ML) classification models trained on the cysteine-liganded co-crystal structures in the protein data bank (PDB) have been reported. The support vector machine (SVM) models trained on 1057 cysteine-liganded co-crystal structures (515 proteins) in the PDB achieved the best AUC (area under the curve of receiver operating characteristic or ROC), recall, and precision of 0.73, 0.62, and 0.41, respectively,[15] in an unseen test. Invariant of the classification threshold, AUC is a primary metric for evaluating the effectiveness of classification models, while recall measures the proportion of actual positives identified correctly, and precision measures the proportion of positive predictions that are actually correct. Most recently, the graph neural network (GNN) models DeepCoSI were trained on the CovalentInDB database which contains 1042 cysteine-liganded co-crystal structures (259 proteins) and achieved the best AUC of 0.92 in the training validation; however, the test metrics are not given.[16]

The arrival of the powerful and continuously improving AlphaFold2 (AF2) structure prediction engine[17] further underscores the potential value of highly predictive structure-based ML models in TCI discovery campaigns. ML models are complementary to the ABPP experiments. For example, the ML model may inform specific cysteine sites that are not easily detectable by chemoproteomics. Coupled to MD simulations, the ML models may also be used to understand how structural or conformational changes, e.g., as a result of protein phosphorylation, impact the cysteine reactivities and ligandabilities, as recently found by ABPP experiments.[18]

Here we report the rigorous development and validation of two types of ML models, the tree-based models and the three-dimensional convolutional neural networks (3D-CNNs), trained on an exhaustively curated new database LigCys3D, comprised of $>$10,000 X-ray crystal structures in the PDB representing 1,133 unique ligandable cysteines in 780 unique proteins. To our best knowledge, this database is the largest to date and significantly surpasses those used for the previous ML models[15,16] in terms of the number of unique proteins and cysteines as well as the number of structural representations. We asked if the current crystal structure information in the PDB is sufficient for developing highly predictive ML models. In multiple tests, our best tree models and CNNs deliver the AUCs of about 94%. Interestingly, testing on the ABPP quantified cysteines that are not in LigCys3D resulted in lowered AUC. We discuss the discrepancies between the liganded cysteines captured by crystallography and those determined by chemoproteomics and the possible ways to improve the models. Our work paves the way for the ML-led in-

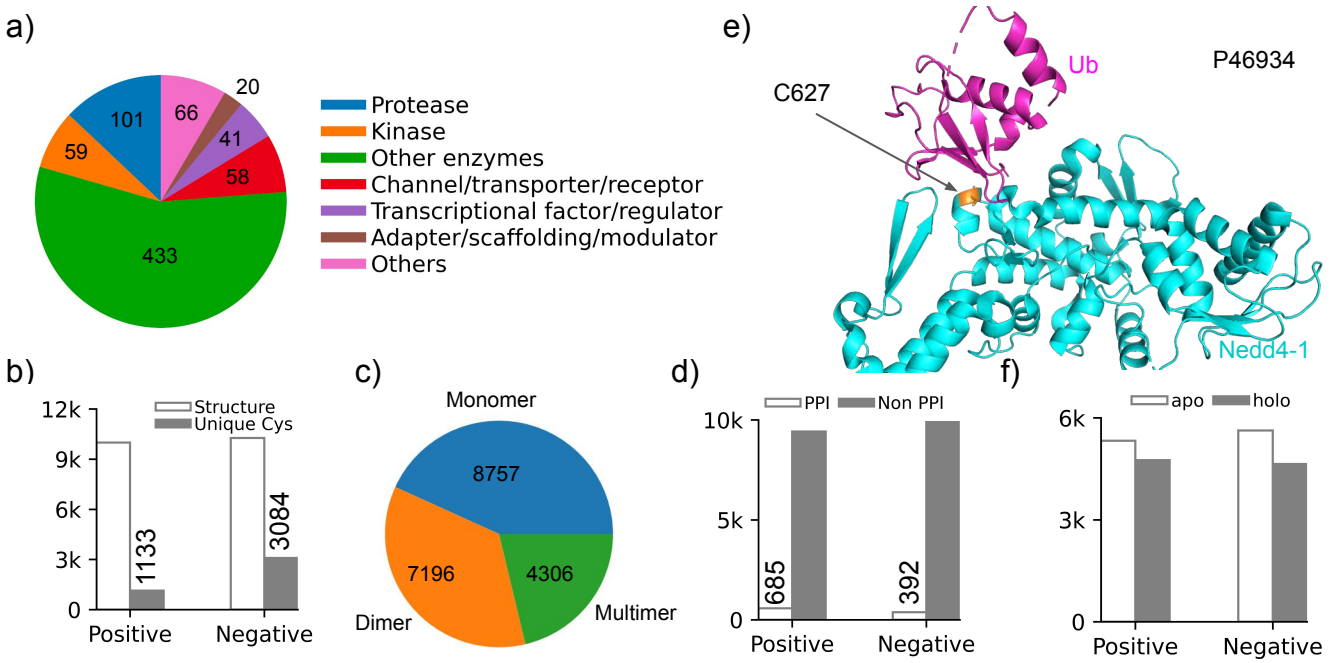

Figure 1. **Analysis of the ligandable cysteines and the associated X-ray structures in the PDB.** a) Functional classes of the proteins that have at least one ligandable (positive) cysteine according to the structures deposited in the PDB. Functional information is taken from the UniProtKB.[19] b) Nedd4-1 (cyan) contains a cysteine (C627, orange) at the PPI with ubiquitin (magenta) in the PDB entry 5C7J. While not liganded in this structure, Cys627 is liganded by a covalent inhibitor in a different, monomeric structure (PDB ID: 5C91). c) Number of unique positive and negative cysteines, and the number of PDB structures containing these cysteines. A positive cysteine is represented by up to 10 PDB structures, and the cysteine is modified in at least one structure. d) Number of (nonunique) cysteines that are in monomer, dimer, and multimer structures based on the biological assembly information in the PDB. e) Number of PDB structures that represent positive or negative cysteines that are near the PPI or not. A PPI cysteine was defined using a distance cutoff of 4.5 Å between the sulfhydrl sulfur and the nearest heavy atom in another chain. f) Number of PDB structures that are apo (ligand free) or holo (bound to any ligand) for positive and negative cysteines.

tegration of big data, structural models, and chemoproteomic experiments to interrogate the proteome space for novel TCI discoveries.

RESULTS and DISCUSSION

Construction of a structure database of cysteine-liganded proteins determined by crystallography. In order to train ML models, we first built a database of proteins containing cysteines that have been covalently modified by ligands. The recently published CovPDB[20] and CovalentInDB[16] databases together contain 659 liganded cysteines in 484 unique proteins. We performed an exhaustive search in the PDB and found additional 474 liganded cysteines in 296 unique proteins. Together, we compiled 1133 liganded cysteines in 780 unique proteins. These cysteines will be referred to as positives. The rest of the 3077 cysteines in these proteins are unliganded, which will be referred to as negatives. We note, although the unliganded cysteines are more reliable negatives than the cysteines in proteins that have not been cysteine-liganded before, false negatives are still possible. Using the most recent PDBx/mmCIF files by SIFTS,[21] we matched each cysteine with the (gene) accession number and residue ID in the UniProt knowledge base (UniProtKB).[19] 76% of the cysteine-liganded proteins are enzymes, including 101 proteases, 59 kinases and 433 other enzymes (Fig. 1a). Channels/transporters/receptors (58), transcription factors and regulators (41) are also present, along with 66 proteins that do not have functional classifications based on UniProtKB[19] or SCOP2[22](Fig. 1a).

The CovPDB[20] and CovalentInDB[16] databases contain only the cysteine-liganded PDB structures, based on which the previously reported ML models were trained.[15,16] This is not ideal, as the conformational variability is neglected, which may limit the model transferability (see later discussion). Thus, we augmented the dataset to a total of 10,105 positive entries (10,105 X-ray structures representing 1,133 positive cysteines) and 97,754 negative entries (97,754 X-ray structures representing 3,084 unique negative cysteines). The quaternary structure was built based on the bioassembly information in the PDB. On average, each positive cysteine is represented by 9 structures, and in most of these structures the positive cysteine is not liganded, i.e., the structure is either ligand free or in complex with a reversible ligand. We will refer to this dataset as LigCys3D (ligandable cysteine three-dimensional structure database) hereafter. Since there are significantly more negatives than positives, we randomly down-sampled the negative entries to 10,267, i.e., 10,267 X-ray structures representing 3,084 negative cysteines (on average, 3 structures per negative cysteine). In total, 20,259 entries were curated as the dataset for model hold-out and training as well as cross-validation (CV, Fig. 1b).

Structural diversity, variability, and allostery are represented in the augmented dataset. Considering the quaternary structures associated with the entries, 8,757 are monomers, 7,196 are dimers, and 4,306 multimers (Fig. 1c). In addition, 685 structures associated with the (119) positive cysteines and 392 structures associated with the (110) negative cysteines are located at the protein-protein (or protein-nucleic acid) interfaces (PPIs, Fig. 1d), as defined using a distance cutoff of 4.5 Å between the cysteine sulfur and its near-

est heavy atom from a different chain in the PDB file. An interesting PPI example is the HECT E3 ubiquitin ligase Nedd4-1, which regulates metabolism, growth, and development and is a promising target for treating cancers and other diseases.[23] Nedd4-1 has a noncatalytic cysteine C627, which is located at the binding interface with ubiquitin (PDB ID: 5C7J)[24] and has been modified by a covalent inhibitor (PDB ID: 5C91).[23] In addition to the cysteine-liganded structures, through data augmentation the positive cysteines are also represented by co-crystal structures in complex with reversible ligands as well as surprisingly more than 50% ligand-free structures. For the positive entries, 3,912 are ligand free and 3,695 are ligand bound, while for the negative entries, 3,601 are ligand free and 3,003 are ligand bound (Fig. 1f). These analyses demonstrate that our data augmentation strategy affords structure diversity and variability, which we surmised to be essential for training truly predictive and transferable models. The inclusion of structural variation may also help with the detection of cryptic pockets.[25] We should also note that in the LigCys3D dataset, each protein has on average 1.5 ligandable cysteines, which suggests that allosteric sites are also represented.

The top three tree models are highly predictive of ligandable cysteines. The recent constant pH MD titration simulations of a large number of kinases uncovered common structural and physical features for reactive cysteines (high tendency to deprotonate at physiological pH) and ligandable cysteines.[12,14,26,27] Thus, we surmised that the feature-based ML classification models such as decision trees may be suited for predicting cysteine ligandabilities. Based on the findings from these studies[12,14,26,27] we devised a set of descriptors (37 after removal of multicollinearity, see Methods) for training the tree-based classifiers using PyCaret.[28] From the down-sampled LigCys3D, 10% of the entries were randomly picked as hold-out for the "unseen" test, while the remaining 90% of the entries were reserved for training/CV. UniProt assession number and residue IDs were used to ensure cysteines are unique between the training/CV and test sets. The 10-fold CV was used, where different folds have unique cysteines. This process (data splitting, training/CV, and test) was repeated six times to generate statistics for model evaluation (Fig. 2a). Following CV, the model was retrained with hyperparameter tuning before being applied to the test set.

The eXtreme Gradient Boosting (XGBoost), Extra Tree (ET), and Light Gradient Boosting (LightGBM) are the top three best performing models according to the AUC, recall, precision, and F1 score in the unseen tests (Table 1). These four metrics analyze the model performance in different ways. The AUC is an aggregate measure of true and false positive rates across all possible classification thresholds. Recall measures the accuracy of the positive predictions given a threshold (percentage of the predicted positives that are truly positive), while precision measures the percentage of positive entries correctly identified. The F1 score is the harmonic mean of recall and precision. Note, we also calculated the selectivity and negative predictive value (NPV), which respectively measure the accuracy and precision of predicting negatives. These metrics are deemphasized in this work because our training set might contain false negatives as discussed before and knowing the positives are more relevant in drug discovery.

The best model XGBoost gave an AUC of 0.94±0.01 (Fig. 2b) and a maximum F1 score of 0.92±0.02, which was achieved at the threshold value of 0.30 (Fig. 2c). With this threshold, the recall and precision are 0.92±0.01 and 0.91±0.01, respectively (Table 1). The test metrics of the ET classifier closely follow those of the XGBoost. Considering the test AUC, recall, and precision of 0.93–0.94, 0.89–0.96, and 0.89–0.91, respectively, the top three tree-based models are highly predictive of ligandable cysteines.

Model performance is unbiased with respect to protein quaternary structure and proximity to interface. It is important to verify that the model performance is unbiased with respect to the protein quaternary structures and proximity to interfaces (if any). We compared the XGBoost model performance metrics for cysteines in the monomer, dimer, and multimer structures (Fig. 2d and Supplemental Table S1). The AUCs for monomers and dimers are identical (0.94) and it is only marginally lower for multimers (0.92). While the recall or precision for monomers and dimers are also identical (0.93 or 0.92, respectively), it is only somewhat lower for multimers (0.87 and 0.86, respectively). As to non-PPI vs. PPI cysteines, the AUC, recall, and precision are nearly identical (Fig. 2e and Supplemental Table S2). These analyses demonstrate that the models are equally predictive for large protein assemblies and PPIs. The latter is desirable, as TCI discovery targeting PPIs has been very challenging.[29]

Cysteine ligandability is determined by a set of structural and physico-chemical features. A significant advantage of decision tree as opposed to neural network models is interpretability. The permutation feature importance scores were calculated to understand the structural and physicochemical features that determine cysteine ligandability. The feature importance score represents the decrease in the model score when a feature is randomly shuffled.[30] Accordingly, cysteine's sidechain solvent-accessible surface area (sasa_side) is by far the most important feature (Fig. 2f), which is readily understood, as solvent exposure promotes cysteine reactivity due to the stabilizing solvation free energy of the anionic thiolate state. However, an earlier study found a poor correlation between the solvent accessibility and thiol reactivity.[31] An early bioinformatics analysis showed that cysteine is the least-exposed amino acid[32] and the recent constant pH MD simulations showed that many hyperreactive cysteines in kinases[26,27] and other proteins[13] are buried. We will come back to this discussion. The next four features: the secondary structure at the cysteine+4 position (dssp_4), the distance from the cysteine sulfur to the nearest pocket (sg_pocket_d1), the distance to the nearest nonpolar atom in another residue (npol_1), and the number of heavy atoms within 15 Å from the cysteine sulfur (n_hv_15), are also consistent with the intuition or knowledge from other studies. In accord with the importance score of dssp_4, the N-terminal capping (Ncap) cysteine on a helix has been suggested as highly reactive two decades ago,[33] which is supported by the fact that the front-pocket Ncap cysteine is the most popular site of targeted covalent inhibition among all kinases.[26] Similar to the BURIED term in the empirical pK_a prediction program PROPKA,[34] the two features npol_1 and n_hv_15 indicate how deeply the cysteine is buried, which affects both the cysteine reactivity and ligand accessibility.

Complementary to the feature importance scores, the game-theoretic SHAP (SHapley Additive exPlanations) val-

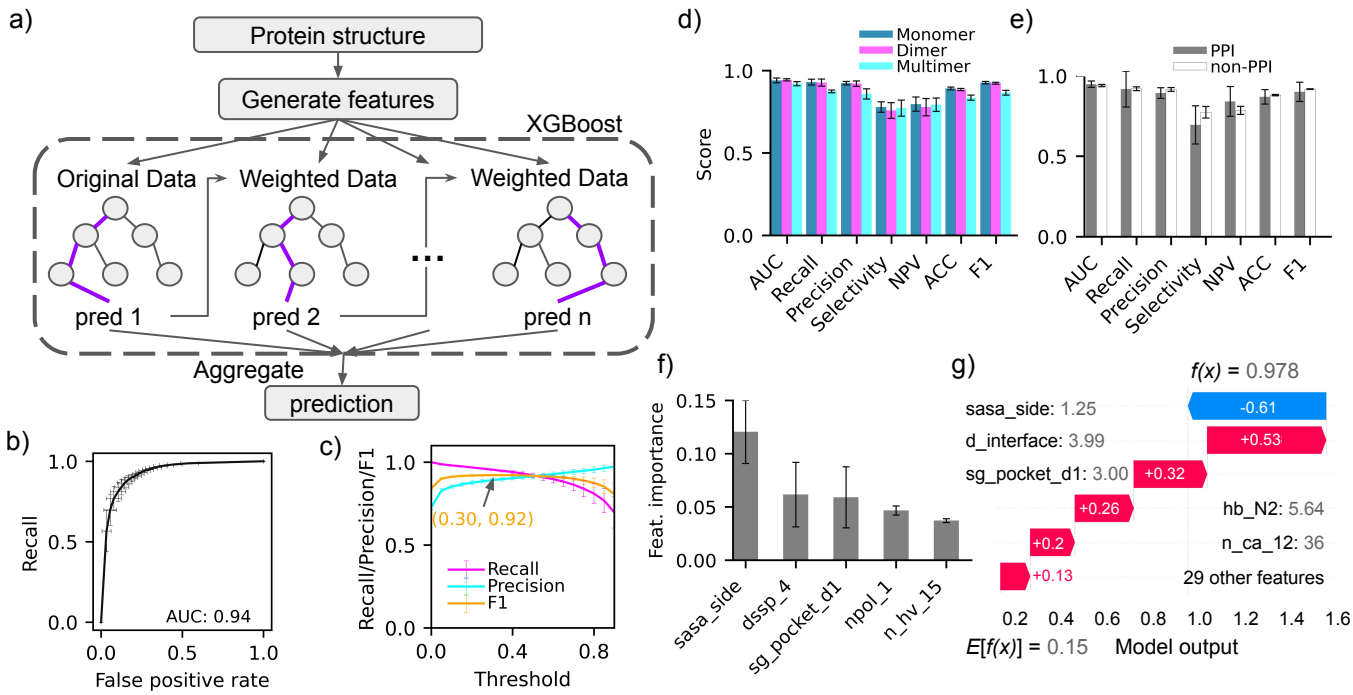

Figure 2. **Performance of the tree-based models for predicting cysteine ligandabilities.** a) Model workflow based on the Extreme Gradient Boosting (XGBoost) classifier. b) Receiver operating characteristic (ROC) curve for the XGBoost models obtained from 6 rounds of data splitting followed by training with 10-fold cross validation. The area under the curve (AUC) is 0.94. c) Recall/Precision/F1 score as a function of the classification threshold. The highest F1 score of 0.92 was achieved at a threshold of 0.30. d) Performance metrics of the ET models for cysteines in monomer, dimer, and multimer structures. e) Performance metrics of the ET models for cysteines at the interfaces (PPIs) or not. f) Permutation-based feature importance scores for the top five features: the sidechain SASA; the distance from the cysteine sulfur to the second nearest nitrogen in His, Asn, Gln, or Trp sidechain (hb_N2), the minimum distance to any nonpolar atom in a different residue (npol_1), and the number of Cα atoms within 12 Å of the cysteine sulfur. g) The "Waterfall" SHAP value plot to explain the ligandability prediction for C627 in Nedd4-1's structure (PDB: 2XBB; UniProt: P46934, C627). The five most impactful features (values are given next to the names) are shown on the top and the rest 29 features are collapsed into one and shown on the bottom. The corresponding SHAP values shown in red (positive) or blue (negative) bars accumulate to shift the expected model output *E[f(x)]* from the random guess output (0.15) to the real output ($f(x) = 0.978$), where *f(x)* is the model output before the logistic link function is applied.

Table 1. Performance of the tree-based and CNN models in the cross validations and unseen tests using the (downsampled) LigCys3D data[a]

Metrics	**ET**		**XGBoost**		**LightGBM**		**CNN**	
	CV	Test	CV	Test	CV	Test	CV	Test
AUC	0.90±0.00	0.94±0.00	0.90±0.00	**0.94±0.01**	0.90±0.00	0.93±0.01	0.98±0.01	**0.93±0.04**
Recall	0.82±0.01	0.89±0.02	0.87±0.01	**0.92±0.01**	0.78±0.01	0.86±0.02	0.93±0.01	**0.96±0.02**
Precision	0.77±0.01	0.93±0.01	0.75±0.01	**0.91±0.01**	0.79±0.01	0.94±0.01	0.92±0.02	**0.89±0.03**
Selectivity	0.81±0.01	0.81±0.02	0.77±0.01	0.77±0.04	0.84±0.01	0.85±0.02	0.94±0.02	0.69±0.10
NPV	0.85±0.00	0.74±0.03	0.89±0.01	0.79±0.03	0.84±0.00	0.70±0.03	0.94±0.01	0.86±0.06
ACC	0.81±0.00	0.87±0.01	0.82±0.00	0.88±0.00	0.82±0.00	0.86±0.01	0.93±0.02	0.88±0.04
F1	0.79±0.00	0.91±0.01	0.80±0.00	**0.92±0.00**	0.78±0.01	0.90±0.01	0.92±0.02	**0.92±0.02**

[a] Metrics are the average and standard deviation from the 10-fold cross validation (CV) or from the external tests by six trained models. Cysteines do not overlap between the training and test datasets. The same train and test sets were used for all models. The test AUC, recall, precision, and F1 score of the top tree model and CNN were highlighted in bold font. The metrics from the null model (random guess) are near 0.5 since the number of positives and negatives is nearly equal in the training and testing sets.

ues inform the impact of feature values on the prediction outcomes.[35,36] A positive or negative SHAP value increases or decreases the model output of a prediction from its expectation value estimated by randomly guessing from the features.[36] As an example, Fig. 2g explains the model prediction for C627 in Nedd4-1 (PDB: 2XBB) based on the SHAP values of the features. While the sasa_side is small (1.25 Å^2) and decreases the model output by 0.61, the other four important features, the cysteine sulfur distance to the interface (d_interface, 3.99 Å), to the nearest pocket (sg_pocket_d1, 3.00 Å), and to the second nearest potential hydrogen bond donor nitrogen (hb_N2, 5.64 Å), as well as the number of Cα atoms within 12 Å of the cysteine sulfur (n_ca_12, 36) increase the model output by 0.53, 0.32, 0.26, and 0.20 respectively. Together with the 0.13 positive contribution from the rest of the features, the model output $f(x)$ is upshifted from the expected value ($E[f(x)]$) of 0.15 to the value of 0.978, which returns a class probability score of 0.73.

The CNN models show similar performance as the XGBoost models. Since many of the tree model features are spatially related, we reasoned that three-dimensional convolutional neural networks (3D-CNN) may offer high performance. We adapted and modified the 3D-CNN architecture of Pafnucy which was developed for protein-ligand binding affinity predictions[37] and recently adapted for protein pK_{a} predictions.[38] In our modified architecture, a cubic grid of 20x20x20 Å with a resolution of 1 Å was created centering at the cysteine sulfur, and each voxel represents a nearby atom and encodes 20 features (Fig. 3a). To remove rotational variance, each cubic box was generated 20 times by randomly rotating the PDB coordinates. The input grid is processed by a block of 3D convolutional layers that have 128 filters (Fig. 3a, details see Methods). To allow comparison to the tree models, data splitting and CV were conducted in the same manner. Interestingly, the 3D-CNN gave very similar to the best tree model XGBoost, with the AUC, accuracy and precision of 0.93$\pm$0.04, 0.96$\pm$0.02, 0.89$\pm$0.03, respectively (Table 1). It is also noteworthy that the standard deviations in the test metrics resulting from the six data splits, training/CV, and testing are overall slightly larger than those of the XGBoost models (Fig. 3b). Although the best average F1 score (0.92) is the same as the XGBoost models, it is achieved with a lower prediction probability threshold (0.15, Fig. 3c).

We also examined the CNN performance for different protein quaternary structures and PPI vs. non-PPI cysteines in comparison to the XGBoost models (Fig. 3d and Table S3). While the AUC, recall, and precision are maintained between monomers and dimers with the XGBoost models, there is a 0.02 decrease in the average AUC or recall and 0.03 decrease in the average precision going from monomers to dimers with the CNN models. As to multimers, the average AUC or recall drop only by 0.01 relative to the dimers (smaller than the XGBoost models) but the precision drops by 0.08 (larger than the XGBoost models). This analysis suggests that the classification power of the CNN models deteriorates slightly more for dimers and multimers as compared to the XGBoost models.

The trend in the model performance differences among quaternary structures is consistent with that between the PPI and non-PPI cysteines (Fig. 3e and Table S4). While the average AUC and recall are maintained going from the non-PPI to the PPI cysteines with the XGBoost models, the respective decrease is 0.03 and 0.02 with the CNNs. As to the precision, the decrease from the non-PPI to the PPI cysteines is only 0.01 as compared to 0.03 with the XGBoost models. Interestingly, the standard deviations among the different CNN tests are doubled going from the non-PPI to the PPI cysteines, which is consistent with the XGBoost tests, although the standard deviations of the latter are overall significantly smaller. One possible reason for the performance deterioration of the CNNs for dimers and multimers is the finite-size grid which may exclude part of the chains that carries relevant information for the model prediction. In contrast, the distance-based features used in the tree models cover all residues in the bioassembly regardless the distances to the cysteine of interest.

Model assessment of the ABPP quantified cysteines in the cells. Given the high performance of the ML models for predicting liganded cysteines captured by crystallography, we asked if the liganded cysteines determined by chemoproteomic experiments in lysate or intact cells but not yet proved by crystallography (i.e., not in LigCys3D) can be recapitulated. For this purpose, we turned to the data from the isotopic tandem orthogonal proteolysis (isoTOP) activity based protein profiling (ABPP) experiments in two cancer cell lines.[10] Here, the cells were treated with electrophilic small-molecule fragments and exposed to the broad-spectrum cysteine reactive probe iodoacetamide (IA)-alkyne; the liganded cysteines were defined as those that showed greater than 75% reduction in IA-alkyne labelling by at least two electrophilic fragments.[10] The overlap between the ABPP quantified (liganded and unliganded) cysteine dataset and LigCys3D is surprisingly small (65), including 21 positive and 44 negative cysteines. To prepare the ABPP test set, we removed these overlapping cysteines, cysteines in disulfides, those inconsistent with other ABPP experiments,[39–43] and those in the proteins that do not have a PDB structure or an Alpha2 (AF2) model representations.[17] We further divided the ABPP data in two sets: the cysteines represented by the PDB structures (ABPP1) and the cysteines only represented by the AF2 models (ABPP2). ABPP1 is comprised of 179 liganded cysteines in 153 proteins and 429 unliganded cysteines in 265 proteins. In the latter category, 237 proteins have no liganded cysteines at all, which is different from LigCys3D where all proteins contain at least one ligandable cysteines.

Table 2 lists the prediction metrics of the three tree models and CNNs for the ABPP1 and ABPP2 datasets along with the null model metrics. The latter represent random guesses, i.e., equal probabilities for predicting positives and negatives. Due to the overwhelming ratio of negatives to positives, the precision is decreased from 0.5 for balanced classes to 0.29 and 0.10 for the ABPP1 and ABPP2 datasets, respectively. For predictions of the ABPP1 data, each cysteine was represented by a single (highest resolution) PDB structure. We focus on the AUC, recall, and precision. Consistent with the unseen tests, the performances of ET, XGBoost, and LightGBM predictions of ABPP1 are on par, with the average AUC of 0.70–0.75, recall of 0.79–0.82, and precision of 0.40–0.44. The latter represents an enrichment of 38%–52% over random guesses, and is equivalent to a precision of 0.69–0.76 given balanced classes. Although lower than the test metrics (above 0.91 for AUC, recall, and precision), these metrics are acceptable given how different crystallography and proteomics experiments are. Surprisingly, the CNNs gave the average AUC, recall, and precision of 0.66, 0.77, and 0.37,

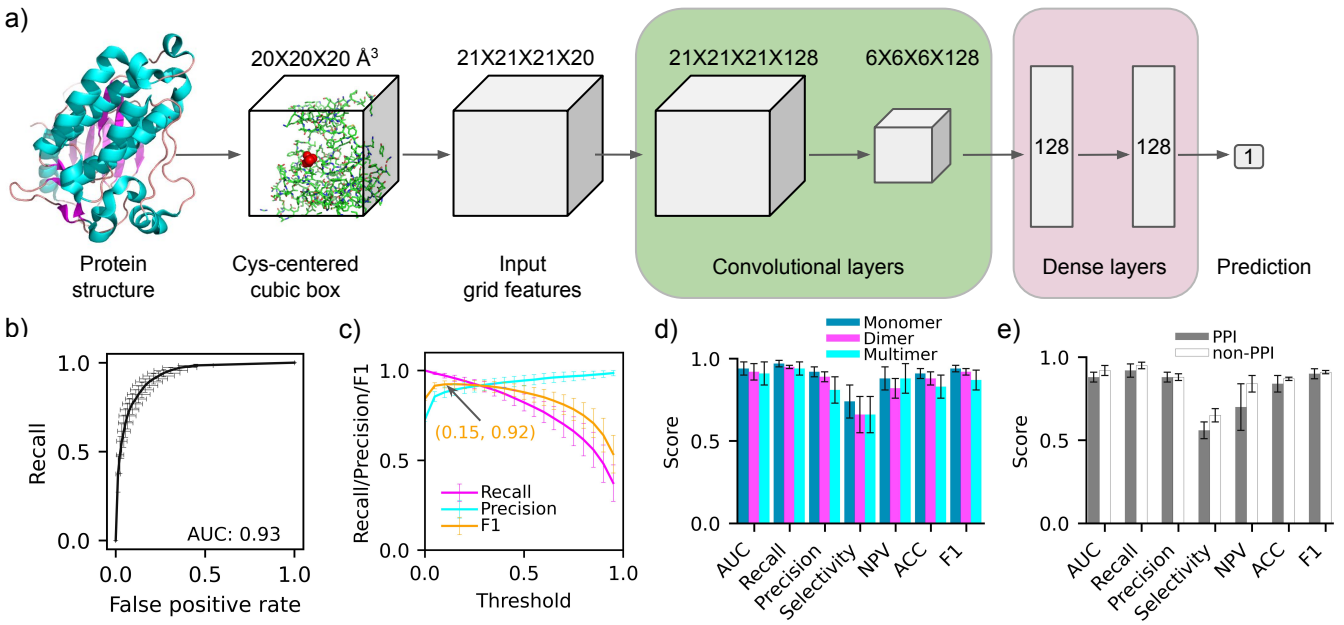

Figure 3. **Performance of the three-dimensional convolutional neural network (CNN).** a) Architecture of the 3D-CNN for cysteine ligandability predictions. b) ROC curve obtained from 6 train/test experiments. The AUC is indicated for the average curve. c) Recall/precision/F1 score as a function of the classification threshold. The best F1 score 0.92 is achieved at a threshold of 0.15. d) Comparison of the CNN performance metrics for cysteines in monomer, dimer, and multimer structures. e) Comparison of the CNN performance metrics for PPI and non-PPI cysteines.

respectively, which are 0.03–0.06 lower than the tree models.

Table 2. Model predictions of the ABPP quantified cysteines that are not in LigCys3D

Metrics	ET	XGBoost	LightGBM	CNN	Null[c]
		ABPP1 (179:429, PDB structures)[a]			
AUC	**0.72±0.01**	0.70±0.01	0.75±0.01	0.66±0.02	0.50
Recall	**0.82±0.02**	0.83±0.01	0.79±0.01	0.77±0.04	0.50
Prec	**0.40±0.00**	0.40±0.01	0.44±0.01	0.37±0.00	0.29
Select	0.49±0.01	0.48±0.01	0.57±0.02	0.45±0.03	0.50
NPV	0.87±0.01	0.87±0.01	0.87±0.00	0.83±0.01	0.71
ACC	0.59±0.00	0.58±0.01	0.64±0.01	0.55±0.01	0.50
F1	0.54±0.01	0.54±0.01	0.56±0.00	0.50±0.01	0.37
		ABPP2 (482:4422, AF2 models)[b]			
AUC	**0.60±0.00**	0.60±0.00	0.60±0.00	0.60±0.01	0.50
Recall	**0.82±0.03**	0.79±0.01	0.72±0.03	0.82±0.04	0.50
Prec	**0.11±0.00**	0.11±0.00	0.12±0.00	0.11±0.00	0.10
Select	0.29±0.04	0.33±0.01	0.41±0.05	0.30±0.04	0.50
NPV	0.94±0.00	0.93±0.00	0.93±0.00	0.94±0.01	0.90
ACC	0.34±0.03	0.38±0.01	0.44±0.04	0.35±0.03	0.50
F1	0.20±0.00	0.20±0.00	0.20±0.00	0.20±0.00	0.17

[a] 179 liganded and 429 unliganded cysteines determined by the isoTOP-ABPP experiments[10] have PDB structures. [b] 509 liganded and 4497 unliganded cysteines have AF2 models only. The average and standard deviation are from the predictions using the 6 trained models. [c] The predictions based on a random decision. The AUC, recall, and precision of the best tree model are in bold font.

For the ABPP2 dataset, since no PDB structures are available, all predictions were made using the AF2 models (Table 2). Although the average AUC and precision of all three tree models are nearly identical, the ET model gave a recall that is 0.03 and 0.10 higher than the XGBoost and LightGBM models, respectively. The ET's average AUC, recall, and precision are 0.60, 0.82, and 0.11, respectively. Note, the recall is the same as in the ABPP1 predictions, however, the precision is much lower and represents only 10% enrichment over random guess (equivalent to 0.55 given balanced classes). As to the CNN models, Since the test performance of the CNN is very similar to that of the best tree model (XGBoost), we expected the prediction metrics for the ABPP data to be similar as well. This is indeed the case for the AF2-based predictions; however, for predictions with PDB structures, the average AUC, recall, and precision of the CNNs are 0.03–0.05 lower than the XGBoost or 0.03–0.06 lower than the ET model (Table 2).

Differences between the liganded cysteines in LigCys3D and ABPP1 datasets. To explain the significant decrease in the AUC, recall, and precision of the models in recapitulating the ABPP quantified cysteines as compared to the unseen tests, we considered several factors. We first compared the LigCys3D and ABPP1 datasets (Fig. 1 and Supplemental Fig. S1). In terms of protein functional classes, enzymes dominate in both datasets. In terms of the percentage of PPI cysteines, it is similar as well. Next, we considered the possibility that some of the unliganded cysteines in our model training set may in fact be ligandable, i.e., negatives are wrongly labeled. However, the distribution of the number of liganded cysteines per protein in LigCys3D is similar to the liganded ABPP1 dataset, i.e., proteins with at least one liganded cysteine (Supplemental Fig. S3). The average number of liganded cysteines per protein in LigCys3D is ∼1.6, as compared to ∼1.2 in the liganded ABPP1 dataset (Supplemental Fig. S3). Thus, the likelihood is low to have significant number of wrongly labeled negatives.

Having ruled out the above, we hypothesized that the the cysteine environment in the protein structures captured by X-ray crystallography may be different from that in the structures representing the ABPP liganded cysteines. To test this hypothesis, we plotted the distributions of the cysteine sidechain SASA and the distance to the nearest pocket, which are important features of the tree models (Fig. 4). For the LigCys3D cysteines, the positives are separated into the modified and unmodified groups, which refer to whether the cysteine is liganded or modified in the structure or not. Note, the unmodified structures can be either apo or in complex

with a reversible ligand. Surprisingly, the major peak of the SASA distribution for the modified positives is at ∼20 Å^2, while that of the unmodified positives is at ∼5 Å^2, which is much closer to the negatives (Fig. 4a, top plot). This suggests that covalent modification on average perturbs the protein structure so as to increase cysteine's solvent exposure. Structural perturbation by reversible ligands is a well-known phenomenon;[44] however, not much is known about the effect of covalent ligands. Comparison of the SASA distributions between the unmodified positives and the negatives demonstrates that while the cysteines in the positives are overall more exposed to solvent, a significant fraction of them are deeply buried. This is consistent with the notion that cysteine is the least-exposed amino acid[32] and our recent finding that most reactive cysteines in kinases[14,26,27] and other proteins are in fact buried.[13]

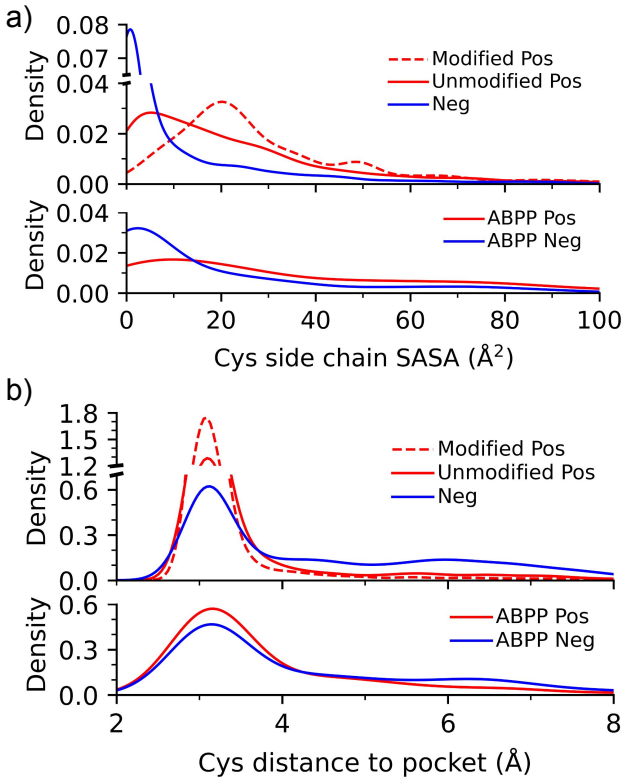

Figure 4. **Comparison of cysteine's structural environment in the LigCys3D (downsampled) and ABPP1 datasets.** a) Distributions of the cysteine sidechain SASA based on the structures in the downsampled LigCys3D (top) and ABPP1 (bottom) datasets. b) Distributions of the distance from the cysteine sulfur to the nearest pocket (alpha sphere) based on the structures in the LigCys3D (downsampled) and ABPP datasets. Modified Pos (dashed red) and unmodified Pos (solid red) refer to the positive structures in which the cysteine is liganded or not.

Now we examine the SASA distributions of the ABPP1 positives and negatives (Fig. 4a, bottom plot). Since the structures representing the ABPP liganded cysteines are not cysteine-liganded, we expected the SASA distribution of the ABPP1 positives to resemble that of the unmodified positives. This is indeed the case, although the ABPP1 distribution is quite flat with a broad maximum in the range of 5–15 Å^2. Similarly, the distribution peak height for the ABPP1 negatives is also much lower than that for the (downsampled) LigCys3D negatives. Overall, the difference in the SASA distribution between the ABPP1 positives and negatives is much smaller than between the LigCys3D positives and negatives.

The distribution of the cysteine distance to the nearest pocket displays a peak near 3 Å for both modified and unmodified positives (near 3 Å, Fig. 4b, top plot); however, the modified positives have a higher peak intensity, suggesting that covalent ligand binding may slightly "pull" the cysteine towards the pocket. Interestingly, the distribution of the negatives also displays a peak near 3 Å, although with a lower peak height as compared to the positives, and importantly, the distribution has a fat tail, suggesting that many negative cysteines are far away from any pocket, as expected. Similar to the SASA distributions, the difference in the pocket distance distribution between the ABPP1 positives and negatives is significantly smaller than the LigCys3D counterparts (Fig. 4b, bottom plot). Furthermore, the pocket distance distribution of ABPP1 positives shows an appreciable population from 4 to 6 Å, which is not the case for the the LigCys3D positives. This suggests that more ABPP1 liganded cysteines are further away from the nearest pocket than those captured by X-ray crystallography.

Training on the cysteine-unliganded structures lowers test performance but improves model transferability. In light of the above finding that ligand modification of a cysteine perturbs the structural environment, we hypothesized that training with the cysteine-unmodified structures alone can improve model prediction of the ABPP1 liganded cysteines. To test this, we retrained the ET models using the modified (model 1), unmodified (model 2), and combined (model 3) structures, and compared their performances in the unseen test and ABPP1 predictions. Surprisingly, the test AUC, recall, and precision of model 1 are all above 0.95, whereas the metrics of model 2 are only between 0.75–0.85 and the model 3 metrics are in-between (Table 3). This may be explained by the observation that the differences in the structure features (e.g., the cysteine SASA) between the modified positives and negatives are much larger than between the unmodified positives and negatives (Fig. 4a, top). Since model 3 uses the combined training data, its test performance is between model 1 and 2.

Model 1 can be directly compared with the published models trained with the cysteine-liganded structures only. The ET metrics (AUC, recall and precision all above 0.95) well surpass the feature-based SVM model (test AUC, recall and precision of 0.73, 0.62, and 0.41),[15] which may be attributed to the larger training set and the physio-chemical features born out of the detailed studies of cysteine reactivities and ligandabilities. The ET model's test AUC (0.96) is also higher than the validation AUC (0.92) of the most recent GNN model (test AUC not known).[16]

In contrast to the test performances, model 1 has by far the lowest recall and model 3 has the highest overall performance in recapitualating the ABPP1 positives (Table 3). Model 3's average AUC, recall, and precision are 0.72, 0.82, and 0.40, respectively. Model 2's average AUC is the same, but the recall and precision is 0.01 lower than model 3. Model 1's AUC is 0.01 lower than model 2 and 3; however, its recall is 0.04 and 0.05 lower than model 2 and 3, respectively. This analysis shows that training on the cysteine-modified structures alone may produce models with "deceptively" high performances but limited transferabilities. The slight increase of the recall and precision in the ABPP1 predictions by model 3 vs. model 2 confirms that data augmentation via structure variation (the number of structures is doubled) enhances

model performance.

Table 3. Impact of training with cysteine-unliganded structures on the ET model predictions of the test and ABPP1 cysteines[a]

Model[b]	Model 1	Model 2	Model 3
Structures	Modified	Unmodified	Combined
Pos:Neg[c]	5931:5931	4061:4061	9992:10267
		Test[d]	
AUC	**0.96±0.00**	0.85±0.02	0.94±0.00
Recall	**0.95±0.01**	0.75±0.03	0.89±0.02
Prec	**0.96±0.01**	0.78±0.03	0.93±0.01
F1	**0.95±0.01**	0.77±0.03	0.91±0.01
		ABPP1 (179:429)[e]	
AUC	0.71±0.01	0.72±0.01	**0.72±0.01**
Recall	0.67±0.02	0.81±0.02	**0.82±0.02**
Prec	0.42±0.00	0.39±0.01	**0.40±0.00**
F1	0.52±0.01	0.53±0.01	**0.54±0.01**

[a]Average and standard deviation of the metrics from the six model predictions are given. The metrics of the best model are highlighted in bold font. [b]Model 1, Model 2, and Model 3 refer to the ET models trained with the cysteine-liganded, cysteine-unliganded, and combined structures, respectively. [c]The number of positives and negatives in the entire dataset (training, CV, and unseen test). [d]The null model metrics for the test are ∼0.5. [e]The ABPP1 dataset is unbalanced (179 positives and 429 negatives). The precision and F1 score of the null model are 0.29 and 0.37, respectively.

Voting-the-best scheme substantially improves recall without sacrificing precision in recapitulating the ABPP1 cysteines. In the model prediction of the ABPP quantified cysteines reported in Table 2 and 3, only one structure was used for each cysteine. An obvious question is how the recall and precision are affected by the use of multiple target structures. By increasing the structural representations of the cysteines of interest, the model performance may also be more rigorously assessed. To address this question, we selected the ABPP1 cysteines with at least two PDB structures, which resulted in 112 positive and 43 negatives cysteines represented by 1486 and 265 structures, respectively. By voting the best, i.e., using the highest predicted class probability among all structures for a given cysteine, the average recall is increased to 0.91 from 0.83 while precision is increased to 0.80 from 0.79, as compared to the single-structure-based predictions (Table 4). This suggests that the increase in true positives somewhat outweighs the additional false positives when incorporating structural variation. As a result, the average F1 score is improved to 0.85 from 0.81 (Table 4). A similar trend can be seen for the CNN predictions by voting the best, where the recall is increased to 0.95 from 0.80, although the precision is slightly lowered by 0.02 as compared to the single-structure-based predictions (Table 4).

To take a closer look at the single-structure vs. voting-the-best scheme, we examined the ET predictions of the ABPP cysteines with the largest number of structure representations: 18 liganded cysteines with at least 20 PDB structures each and 14 unliganded cysteines with at least 5 PDB structures each (Fig. 5). The single scheme gave 14 true positives (TPs), 4 false negatives (FNs), and 8 false positives (FPs), resulting in a recall of 78% and a precision of 64%. Using the voting-the-best scheme, the TPs increase to 18, the FNs decrease to 0, and the FPs increase to 9, which results in a recall of 100% and a precision of 61%. In this example, voting-the-best scheme significantly improves the recall and only marginally worsens the precision.

We also tested the voting-by-majority scheme, i.e., classification by the majority of structures. In the ET predictions, the AUC, recall, precision, and F1 remain identical to the single-structure predictions, while for the CNNs, the AUC and recall are increased by 0.01 and 0.04, respectively, and precision remains the same (Table 4). Together, these results demonstrate that sampling diverse structural representations of a cysteine improves the recall of ligandable cysteines without sacrificing the precision and voting-the-best scheme outperforms voting-by-majority scheme.

Table 4. Impact of using multiple target structures on the ET and CNN model predictions of the ABPP1 cysteines[a]

Metrics	Best[b]	Major[c]	Single[d]	Null[e]
		ET		
AUC	**0.70±0.01**	0.70±0.01	0.71±0.01	0.50
Recall	**0.91±0.01**	0.84±0.02	0.83±0.02	0.50
Prec	**0.80±0.00**	0.79±0.00	0.79±0.00	0.72
F1	**0.85±0.00**	0.81±0.01	0.81±0.01	0.59
		CNN		
AUC	**0.68±0.01**	0.69±0.01	0.68±0.01	0.50
Recall	**0.95±0.01**	0.81±0.03	0.80±0.04	0.50
Prec	**0.78±0.01**	0.80±0.01	0.80±0.01	0.72
F1	**0.86±0.01**	0.80±0.02	0.80±0.02	0.59

[a]The ABPP1 cysteines with at least two PDB structures. [b]Classification using the highest predicted class probability for all structures. [c]Classification by the majority (≥50%) of structures. [d]Predictions based on a single (highest resolution) crystal structure. [e]The null-model metrics represent random guesses from 112 positives and 43 negatives.

Why is it challenging to recapitulate the ABPP ligandability data without PDB structures? Recapitulating the ABPP quantified cysteines without PDB structures (ABPP2 dataset) proved challenging. The AUC is decreased from 0.72 in the prediction of ABPP1 cysteines to 0.60 using all three tree models and CNNs. Overall, the performances are similarly poor across all models. We hypothesized that the structure representations by the AF2 models may not be accurate. To test this, we separated out the ABPP2 cysteines represented by the high quality AF2 models, as defined here by the average per-residue confidence score pLDDT (predicted local distance difference test)[17,45] $>$ 90 greater than 90 for the entire chain and for the cysteine of interest. The prediction AUC for this small subset of cysteines (71 positives and 960 negatives) remains the same (Supplemental Table S6), suggesting that the model accuracy is not the culprit.

We next considered protein classes, as enzymes dominate the LigCys3D and ABPP1 datasets (Fig. 1 and Supplemental Fig. S1), whereas the majority of proteins in the ABPP2 dataset are not found in the SCOP2 database,[22] Supplemental Fig. S2), suggesting that they have unknown functional classes. To test if the models indeed have more predictive power for enzymes, the ET models were used to predict the cysteines in the ABPP2 enzymes (44 positives and 904 negatives, Supplemental Table S7). The AUC is increased from 0.60 in predicting all ABPP2 cysteines to 0.66, demonstrating that the classification power of the models is indeed strong for enzymes albeit to a small degree. One possibility is that among the proteins with unknown functional classes are a large number of transmembrane proteins which are underrepresented in the model training set. It is also possible that lipid-facing cysteines in transmembrane proteins may be false positives, as they would be calculated as solvent exposed due to the lack of membrane representation. Other

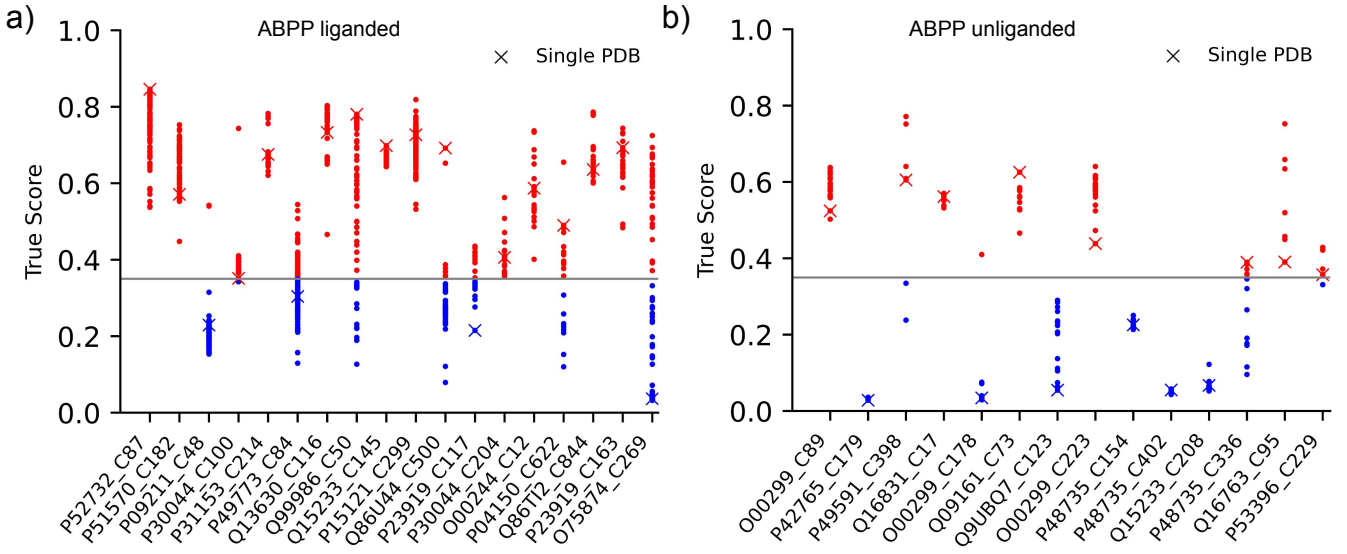

Figure 5. **ET predictions of the ABPP1 cysteines using the single-structure and voting-the-best schemes.** a) Predicted class probabilities of the 18 ABPP1 liganded cysteines (labeled by the UniProt assession numbers) that have at least 20 PDB structures each. b) Predicted class probabilities of the 14 ABPP1 unliganded cysteines that have at least 5 PDB structures each. Every structure was used to make a prediction. The classification threshold is indicated by the grey line. The positive and negative predictions are colored red and blue, respectively. The class probabilities in the single-structure scheme (same as in Table 2) are marked as crosses. One ET model was used for making predictions.

factors that may contribute to the lower AUC in classifying the ABPP2 cysteines include the lack of information of the protein quaternary structure and more importantly, the bound ligand or interaction partner in a complex.

CONCLUDING DISCUSSION

Exploiting a newly curated comprehensive database (LigCys3D) of liganded cysteines captured by X-ray crystallography, we have developed the tree-based and 3D-CNN models for cysteine ligandability predictions. In multiple unseen tests, the ET and XGBoost models gave the AUC of 94%, while the CNN models gave the AUC of 93%. An initially surprising finding is that the models trained on the cysteine-liganded structures only (model 1) have extremely high performances (AUC of 96%), as compared to models trained on the apo and reversible ligand bound structures. This can be explained by the significant difference in the structural environment (e.g., solvent exposure and distance to the nearest pocket) between the ligandable and unligandable cysteines in the cysteine-liganded structures, and the difference is much smaller in the cysteine-unliganded structures. In other words, covalent labeling perturbs the protein structure, making the classification task easier. As expected, model 1 recall drastically decreases (to 67%) when applied to the ABPP1 data, demonstrating the limited transferability. In contrast, models trained on the cysteine-unliganded or combined structures give the recall of 81% or 82%, demonstrating that accounting for structure variability is critical for developing transferable models.

Another surprising finding is that with only 37 structural and physico-chemical features, the top three tree-based models (e.g., AUC of 72% for ET) outperform the CNNs (AUC of 66%) in recapitulating the ABPP cysteines with PDB structures, although the test performances are on par. This supports the notion that incorporating physical properties can improve model transferability. It is encouraging to see that with voting-the-best scheme significantly increases the recall without sacrificing precision in predicting the ABPP1 cysteines. The ET models using the voting-the-best scheme achieved the recall of 91% as compared to 83% using the single-structure scheme.

Recapitulation of the ABPP2 ligandability data without PDB structures proved challenging; the AUC is lowered to 60% from 72% (single-structure) with PDB structures. Unlike the ABPP1 dataset where enzymes dominate, ABPP2 dataset contains mostly proteins for which the functional classes have not been annotated by SCOP2.[22] A significant fraction of them may be transmembrane proteins for which the lipid environment is not taken into account by the models leading to possible false positives (e.g. lipid-facing cysteines). Missing biological details such as the presences of ligand, cofactors, protein quaternary structure and binding partners may also contribute to the discrepancies between model predictions and cell-based ABPP data. These are issues that can be addressed in the future using emerging tools derived from AlphaFold2,[17] e.g., to generate heterodimeric protein complexes[46] or to add bound ligand and cofactors.[47] However, the lack of conformational variability and more importantly, the model representation of the flexible and often functionally important regions remains a weakness for AlphaFold2 models.[48]

The model assessment of the ABPP quantified cysteines led us to contemplate the divergence between the ligandabilities captured by crystallography and those determined by ABPP. First, the liganded cysteines in transmembrane proteins and other non-enzyme protein families are significantly underrepresented in the PDB but they dominate in the cell-based ABPP experiments. A significant fraction of ABPP quantified proteins have yet to be annotated functional classes. Second, the difference between the local structures of the liganded and unliganded cysteines in the ABPP dataset is much more subtle as compared to the LigCys3D dataset. This observation has several implications. As the ABPP experiments are conducted in cells,[10] some cysteines may be ligandable only in the cellular environment, e.g., due to the interaction with another protein as suggested in a recent experiment[49] or due to a post-translational modifica-

tion (e.g., phosphorylation) that may increase the cysteine reactivity as demonstrated recently.[18] Conformational state and/or cysteine environment may also be influenced by the binding of endogenous ligand or cofactor in the cells, which in turn modulates cysteine ligandability. Additionally, mutations may also perturb the structure and cysteine ligandability, but mutants and wild types are not differentiated by our models or in the published ABPP datasets[10] due to the use of UniProt assession numbers for protein identification. This may be exacerbated given that cancer cell lines were used in the ABPP experiment[10] against which our models are tested.

Without the knowledge of the membrane environment, post-translational modification, binding and interaction partner, or mutation, training models on the ABPP data is challenging, as evidenced by the significant lower test AUC (77%) of a preliminary ET model trained on the ABPP1 data, as compared to the ET models trained on the cysteine-unliganded structures alone (AUC of 85%). In light of the above considerations, future work will be directed at incorporating biological context and the optimum use of AlphaFold2 models as well as combining MD simulations to account for state dependence. Developing ML models as a surrogate of crystallography may also further unleash the power of chemoproteomics, accelerating the discoveries of first-in-class therapeutics. Our work represents a first step towards the ML-led integration of big genome data, structure models, and chemoproteomics experiments to annotate the human proteome for the next-generation drug discoveries.

Materials and Methods

Construction of the LigCys3D database. Two recently published databases, CovPDB[20] and CovalentInDB,[16] compiled cysteine-liganded co-crystal structures in the RCSB Protein Data bank (PDB). These two databases have overlap and together they provide 2875 cysteine-liganded co-crystal structures representing 662 liganded cysteines in 489 unique proteins. We conducted an exhaustive search in the PDB and found additionally 472 liganded cysteines in 294 unique proteins. We note, the "L-peptide linking"[50] cysteines that were chemically modified at locations other than the sulfur (SG) atom or simply oxidized were excluded, as well as the cysteines involved in disulfide bonds, zinc-finger coordination, or iron-sulfur clusters. Following the compilation of the cysteine-liganded structures, we used SIFTS[21] to annotate the liganded cysteines with UniProt assession numbers and residue IDs (`https://www.uniprot.org`),[19] which allowed us to retrieve all PDB entries associated with these cysteines. We refer to a cysteine as positive if it is liganded in any crystal structure, and the other cysteines in these structures are referred to as negatives. Note, the bioassembly structures (CIF files) were downloaded, and the coordinates of missing atoms or residues if any were added using pdbfixer (https://github.com/openmm/pdbfixer).[51] We refer to this dataset as LigCys3D.

Data engineering for the ML models. To constructed a ML training set with balanced positive and negative classes and to reduce model training time, we down-sampled the number of structures in LigCys3D as follows. For each unique positive cysteine (based on the UniProt assession number and residue ID), all cysteine-liganded structures were kept and up to four cysteine-unliganded structures were selected (see below). We refer to these structures as the liganded and unliganded positives, respectively. For each unique negative cysteine, up to ten structures were selected (see below). To maximize structural variation, the unliganded positive structures were put into four bins based on the cysteine sidechain solvent accessible surface area (SASA) values, and one structure was randomly picked from each bin. Similarly, the structures representing a negative cysteine were put into ten bins based on the SASA values, and one structure was randomly picked from each bin. Subsequently, a cysteine ligandability data set (down-sampled from LigCys3D) was constructed, comprising 9,992 positive (1,133 unique positive cysteines in 9,992 structures) and 10,267 negative (3,084 unique negative cysteines in 10,267 structures) entries. We will use this data set for model training and testing.

Feature engineering for the tree models. Features are critical for the performance of tree-based models. We conceived a set of structural and physico-chemical features based on our findings from the constant pH molecular dynamics (MD) analysis of cysteine reactivities and ligandabilities in a large number of kinases[12,14,26,27] and other enzymes.[13] In total, eights types of features were calculated based on the input structure, including solvent accessibility (proximity to hydrophobic residues and the cysteine SASA calculated with NACCESS[52]); distance to pockets (defined by fpocket[53]); potential hydrogen bonding; electrostatic interactions; secondary structures (calculated with Biopython[54]); residue flexibility (calculated with PredyFlexy[55]); distance to protein-protein/nucleic acid interface; and presence or absence of ligand binding. A detailed list of the features that were tested is given in Supplemental Methods. After removing highly correlated features, 37 features were left (see Supplemental Methods).

Training of the tree-based models. PyCaret[28] was used for building tree-based classifiers. We manually separated 10% of the data as the unseen test set and 90% as training set (see below). The training set was used for the 10-fold cross validations (CVs). To ensure that the training and testing sets do not contain structures representing the same cysteine, we first grouped the structures according to the UniProt residue IDs and then performed the training-test split by the UniProt residue IDs. In the cross validation, the group-Kfold method was used to avoid putting structures associated the same cysteine in different folds. Multicollinearity was removed with a threshold of 0.9. This leads to a total of 37 features (see above). Categorical features were one-hot encoded. Model training used the binary cross-entropy as loss function and default hyperparameters. The default scikit-learn search library was used to search the hyper-parameters, which were tuned using the tune_model function in PyCaret 5000 times by optimizing the F1 score across all validation folds. Following tuning, the best hyper-parameters were used to train the entire training set, and the final model was saved for predictions on the unseen test set or the ABPP data set. Feature importance scores were generated using the evaluate_model function. To generate statistics for model evaluation, the above process was repeated 6 times, and the average and standard deviation of the model performance metrics were calculated.

Training of the convolutional neural networks (CNNs). The test-train splitting and 10-fold cross validation were performed in the same manner as the tree models. The 3D-CNN architecture was adapted and modified from the Pafnucy model,[37] which was recently adapted for protein pK_a predictions.[38] The input of the CNN represents a 3D image of the protein with 20 color channels. Specifically, a 20-Å 3D grid centered at the SG atom of the cysteine of interest was created. The protein heavy atoms were mapped to the grid with a 1-Å resolution, and each grid point was encoded with 20 features (the default is zero if no atoms): one-hot encoding of 5 atom types C, N, O, S, and others; 1 integer (1/2/3) for atom hybridization; 1 integer for the number of bonded heavy atoms; 1 integer for the number of bonded hetero atoms; one-hot encoding (5 in total) of the SMARTS patterns[56] hydrophobic, aromatic, acceptor, donor, and ring; 1 float for grid charge; one-hot encoding of 6 residue types Asp/Glu, Lys/Arg, His, Cys, Asn/Gln/Trp/Tyr/Ser/Thr, and others. Each cubic box was generated 20 times by rotating the coordinates in the PDB structure to remove rotational variance.

Keras[57] was used to build the CNN. The CNN model contains two Conv3D layers and each Conv3D layer has 128 filters, kernel size 5, activation function relu, and 'same' padding, followed by a pool size 2 MaxPool3D layer and a BatchNormalization layer. Next, a GlobalAveragePooling3D layer is added to do global pooling and then the data is flatterned by a 128 units Dense layer with relu activation, normalized by a BatcheNormalization layer, and filtered by a 0.5 ratio Dropout layer. Finally, a Dense layer of 1 unit and sigmoid activation function is used to generate a binary classification result. Batch size is set to 32 and binary cross-entropy is used as loss function. 50 epochs of training in Adam optimizer is used, with the learning rate of 0.0001 and early stopping if the validation loss plateaus in 5 epochs. The model with the lowest loss in the validation set is saved for the tests on the unseen LigCys3D data and the ABPP data. In these tests, we used the voting result based on the predictions by the 10 saved models from CVs. The voting threshold was determined by the average F1 score on the test set across 6 train/CV:test splitting experiments.

Tests on the ABPP quantified cysteines that are not been in LigCys3D. The ABPP experiment of Backus et al.[10] identified 758 liganded and 5399 unliganded cysteines. After removing the 36 positive and 31 negative cysteines that are in LigCys3D and the negatives that were found positive by other experiments,[39–43] as well as those that do not have PDB structures or AlphaFold2 models,[17] 190 positive and 439 negative cysteines were left, which were used in the predictions. X-ray structures with the highest resolution for the positive and negative cysteines were fetched from the RCSB. AlphaFold2 models for the same residues available in RCSB PDB were downloaded from the European Bioinformatics Institute website https://alphafold.ebi.ac.uk/.[17,58] Other treatments were the same as in the test predictions.

Calculation of the model performance metrics. Given a confusion matrix comprised of the number of true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs), the model performance metrics, recall (or true positive rate TPR), precision, specificity, negative predictive value (NPV), accuracy (ACC), and F1 score are defined as follows.

$$\text{Recall} = TP/(TP + FN) \quad (1)$$

$$\text{Precision} = TP/(TP + FP) \quad (2)$$

$$\text{Selectivity} = TN/(TN + FP) \quad (3)$$

$$\text{NPV} = TN/(TN + FN) \quad (4)$$

$$\text{ACC} = (TP + TN)/(TP + TN + FP + FN) \quad (5)$$

$$F1 = 2 * \text{Recall} * \text{Precision}/(\text{Recall} + \text{Precision}) \quad (6)$$

AUC is calculated by integrating the area under the ROC (receiver operating characteristic) curve, which consists of the recall and false positive rate (1 - selectivity) at all possible classification threshold values.

Supporting Information Available Supporting Information contains supplemental methods, tables, and figures. All training, testing, and validation on the ABPP data as well as models are downloadable from `https://github.com/JanaShenLab/DeepCys`.

Acknowledgement We acknowledge financial support by the National Cancer Institute (R01CA256557).

References

(1) Bauer, R. A. Covalent Inhibitors in Drug Discovery: From Accidental Discoveries to Avoided Liabilities and Designed Therapies. *Drug Discov. Today* **2015**, *20*, 1061–1073.
(2) Gehringer, M.; Laufer, S. A. Emerging and Re-Emerging Warheads for Targeted Covalent Inhibitors: Applications in Medicinal Chemistry and Chemical Biology. *J. Med. Chem.* **2019**, *62*, 5673–5724.
(3) Lu, W.; Kostic, M.; Zhang, T.; Che, J.; Patricelli, M. P.; Jones, L. H.; Chouchani, E. T.; Gray, N. S. Fragment-Based Covalent Ligand Discovery. *RSC Chem. Biol.* **2021**, *2*, 354–367.
(4) Ostrem, J. M.; Peters, U.; Sos, M. L.; Wells, J. A.; Shokat, K. M. K-Ras(G12C) Inhibitors Allosterically Control GTP Affinity and Effector Interactions. *Nature* **2013**, *503*, 548–551.
(5) Cravatt, B. F.; Wright, A. T.; Kozarich, J. W. Activity-Based Protein Profiling: From Enzyme Chemistry to Proteomic Chemistry. *Annu. Rev. Biochem.* **2008**, *77*, 383–414.
(6) Sanman, L. E.; Bogyo, M. Activity-Based Profiling of Proteases. *Annu. Rev. Biochem.* **2014**, *83*, 249–273.
(7) Cravatt, B. F. Activity-Based Protein Profiling – Finding General Solutions to Specific Problems. *Israel J. Chem.* **2023**,
(8) Boike, L.; Henning, N. J.; Nomura, D. K. Advances in Covalent Drug Discovery. *Nat. Rev. Drug Discov.* **2022**, *21*, 881–898.
(9) Chen, Y.; Craven, G. B.; Kamber, R. A.; Cuesta, A.; Zhersh, S.; Moroz, Y. S.; Bassik, M. C.; Taunton, J. Direct mapping of ligandable tyrosines and lysines in cells with chiral sulfonyl fluoride probes. *Nat. Chem.* **2023**, *xxxx*, xxxx–xxxx.
(10) Backus, K. M.; Correia, B. E.; Lum, K. M.; Forli, S.; Horning, B. D.; González-Páez, G. E.; Chatterjee, S.; Lanning, B. R.; Teijaro, J. R.; Olson, A. J.; Wolan, D. W.; Cravatt, B. F. Proteome-Wide Covalent Ligand Discovery in Native Biological Systems. *Nature* **2016**, *534*, 570–574.
(11) Kuljanin, M.; Mitchell, D. C.; Schweppe, D. K.; Gikandi, A. S.; Nusinow, D. P.; Bulloch, N. J.; Vinogradova, E. V.; Wilson, D. L.; Kool, E. T.; Mancias, J. D.; Cravatt, B. F.; Gygi, S. P. Reimagining High-Throughput Profiling of Reactive Cysteines for Cell-Based Screening of Large Electrophile Libraries. *Nat. Biotechnol.* **2021**, *39*, 630–641.
(12) Liu, R.; Yue, Z.; Tsai, C.-C.; Shen, J. Assessing Lysine and Cysteine Reactivities for Designing Targeted Covalent Kinase Inhibitors. *J. Am. Chem. Soc.* **2019**, *141*, 6553–6560.
(13) Harris, R. C.; Liu, R.; Shen, J. Predicting Reactive Cysteines with Implicit-Solvent-Based Continuous Constant pH Molecular Dynamics in Amber. *J. Chem. Theory Comput.* **2020**, *16*, 3689–3698.
(14) Romany, A.; Liu, R.; Zhan, S.; Clayton, J.; Shen, J. Analysis of the ERK Pathway Cysteinome for Targeted Covalent Inhibition of RAF and MEK Kinases. *J. Chem. Inf. Model.* **2023**, *63*, 2483–2494.
(15) Zhang, W.; Pei, J.; Lai, L. Statistical Analysis and Prediction of Covalent Ligand Targeted Cysteine Residues. *J. Chem. Inf. Model.* **2017**, *57*, 1453–1460.
(16) Du, H.; Jiang, D.; Gao, J.; Zhang, X.; Jiang, L.; Zeng, Y.; Wu, Z.; Shen, C.; Xu, L.; Cao, D.; Hou, T.; Pan, P. Proteome-Wide Profiling of the Covalent-Druggable Cysteines with a Structure-Based Deep Graph Learning Network. *Research* **2022**, *2022*, 9873564.
(17) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.;

Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.
(18) Kemper, E. K.; Zhang, Y.; Dix, M. M.; Cravatt, B. F. Global Profiling of Phosphorylation-Dependent Changes in Cysteine Reactivity. *Nat. Methods* **2022**, *19*, 341–352.
(19) The UniProt Consortium,; Bateman, A.; Martin, M.-J.; Orchard, S.; Magrane, M.; Agivetova, R.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E. H.; Britto, R.; Bursteinas, B.; Bye-A-Jee, H.; Coetzee, R.; Cukura, A.; Da Silva, A.; Denny, P.; Dogan, T.; Ebenezer, T.; Fan, J.; Castro, L. G.; Garmiri, P.; Georghiou, G.; Gonzales, L.; Hatton-Ellis, E.; Hussein, A.; Ignatchenko, A.; Insana, G.; Ishtiaq, R.; Jokinen, P.; Joshi, V.; Jyothi, D.; Lock, A.; Lopez, R.; Luciani, A.; Luo, J.; Lussi, Y.; MacDougall, A.; Madeira, F.; Mahmoudy, M.; Menchi, M.; Mishra, A.; Moulang, K.; Nightingale, A.; Oliveira, C. S.; Pundir, S.; Qi, G.; Raj, S.; Rice, D.; Lopez, M. R.; Saidi, R.; Sampson, J.; Sawford, T.; Speretta, E.; Turner, E.; Tyagi, N.; Vasudev, P.; Volynkin, V.; Warner, K.; Watkins, X.; Zaru, R.; Zellner, H.; Bridge, A.; Poux, S.; Redaschi, N.; Aimo, L.; Argoud-Puy, G.; Auchincloss, A.; Axelsen, K.; Bansal, P.; Baratin, D.; Blatter, M.-C.; Bolleman, J.; Boutet, E.; Breuza, L.; Casals-Casas, C.; De Castro, E.; Echioukh, K. C.; Coudert, E.; Cuche, B.; Doche, M.; Dornevil, D.; Estreicher, A.; Famiglietti, M. L.; Feuermann, M.; Gasteiger, E.; Gehant, S.; Gerritsen, V.; Gos, A.; Gruaz-Gumowski, N.; Hinz, U.; Hulo, C.; Hyka-Nouspikel, N.; Jungo, F.; Keller, G.; Kerhornou, A.; Lara, V.; Le Mercier, P.; Lieberherr, D.; Lombardot, T.; Martin, X. et al. UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480–D489.
(20) Gao, M.; Moumbock, A. F. A.; Qaseem, A.; Xu, Q.; Günther, S. CovPDB: A High-Resolution Coverage of the Covalent Protein–Ligand Interactome. *Nucleic Acids Res.* **2022**, *50*, D445–D450.
(21) Choudhary, P.; Anyango, S.; Berrisford, J.; Tolchard, J.; Varadi, M.; Velankar, S. Unified Access to Up-to-Date Residue-Level Annotations from UniProtKB and Other Biological Databases for PDB Data. *Sci. Data* **2023**, *10*, 204.
(22) Andreeva, A.; Kulesha, E.; Gough, J.; Murzin, A. G. The SCOP Database in 2020: Expanded Classification of Representative Family and Superfamily Domains of Known Protein Structures. *Nucleic Acids Res.* **2020**, *48*, D376–D382.
(23) Kathman, S. G.; Span, I.; Smith, A. T.; Xu, Z.; Zhan, J.; Rosenzweig, A. C.; Statsyuk, A. V. A Small Molecule That Switches a Ubiquitin Ligase From a Processive to a Distributive Enzymatic Mechanism. *J. Am. Chem. Soc.* **2015**, *137*, 12442–12445.
(24) Zhang, W.; Wu, K.-P.; Sartori, M. A.; Kamadurai, H. B.; Ordureau, A.; Jiang, C.; Mercredi, P. Y.; Murchie, R.; Hu, J.; Persaud, A.; Mukherjee, M.; Li, N.; Doye, A.; Walker, J. R.; Sheng, Y.; Hao, Z.; Li, Y.; Brown, K. R.; Lemichez, E.; Chen, J.; Tong, Y.; Harper, J. W.; Moffat, J.; Rotin, D.; Schulman, B. A.; Sidhu, S. S. System-Wide Modulation of HECT E3 Ligases with Selective Ubiquitin Variant Probes. *Mol. Cell* **2016**, *62*, 121–136.
(25) Oleinikovas, V.; Saladino, G.; Cossins, B. P.; Gervasio, F. L. Understanding Cryptic Pocket Formation in Protein Targets by Enhanced Sampling Simulations. *J. Am. Chem. Soc.* **2016**, *138*, 14257–14263.
(26) Liu, R.; Zhan, S.; Che, Y.; Shen, J. Reactivities of the Front Pocket N-Terminal Cap Cysteines in Human Kinases. *J. Med. Chem.* **2022**, *65*, 1525–1535.
(27) Liu, R.; Verma, N.; Henderson, J. A.; Zhan, S.; Shen, J. Profiling MAP Kinase Cysteines for Targeted Covalent Inhibitor Design. *RSC Med. Chem.* **2022**, *13*, 54–63.
(28) Ali, M. PyCaret: An Open Source, Low-Code Machine Learning Library in Python. 2020.
(29) Lu, H.; Zhou, Q.; He, J.; Jiang, Z.; Peng, C.; Tong, R.; Shi, J. Recent Advances in the Development of Protein–Protein Interactions Modulators: Mechanisms and Clinical Trials. *Sig. Transduct. Target Ther.* **2020**, *5*, 213.
(30) Statistics, L. B.; Breiman, L. Random Forests. Mach. Learn. 2001; pp 5–32.
(31) Junutula, J. R.; Bhakta, S.; Raab, H.; Ervin, K. E.; Eigenbrot, C.; Vandlen, R.; Scheller, R. H.; Lowman, H. B. Rapid identification of reactive cysteine residues for site-specific labeling of antibody-Fabs. *J. Immun. Methods* **2008**, *332*, 41–52.
(32) Marino, S. M. Cysteine Function Governs Its Conservation and Degeneration and Restricts Its Utilization on Protein Surfaces. *J. Mol. Biol.* **2010**, *404*, 902–916.
(33) Anderson, T. A.; Sauer, R. T. Role of an Ncap Residue in Determining the Stability and Operator-Binding Affinity of Arc Repressor. *Biophys. Chem.* **2002**, *100*, 341–350.
(34) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical p K_a Predictions. *J. Chem. Theory Comput.* **2011**, *7*, 525–537.
(35) Shapley, L. S. *A Value for n-Person Games*; RAND Corporation, 1952.
(36) Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. Adv. Neural Inf. Process. Syst. 2017.
(37) Stepniewska-Dziubinska, M. M.; Zielenkiewicz, P.; Siedlecki, P. Development and Evaluation of a Deep Learning Model for Protein–Ligand Binding Affinity Prediction. *Bioinformatics* **2018**, *34*, 3666–3674.
(38) Cai, Z.; Luo, F.; Wang, Y.; Li, E.; Huang, Y. Protein p K_a Prediction with Machine Learning. *ACS Omega* **2021**, *6*, 34823–34831.
(39) Vinogradova, E. V.; Zhang, X.; Remillard, D.; Lazar, D. C.; Suciu, R. M.; Wang, Y.; Bianco, G.; Yamashita, Y.; Crowley, V. M.; Schafroth, M. A.; Yokoyama, M.; Konrad, D. B.; Lum, K. M.; Simon, G. M.; Kemper, E. K.; Lazear, M. R.; Yin, S.; Blewett, M. M.; Dix, M. M.; Nguyen, N.; Shokhirev, M. N.; Chin, E. N.; Lairson, L. L.; Melillo, B.; Schreiber, S. L.; Forli, S.; Teijaro, J. R.; Cravatt, B. F. An Activity-Guided Map of Electrophile-Cysteine Interactions in Primary Human T Cells. *Cell* **2020**, *182*, 1009–1026.e29.
(40) Yan, T.; Desai, H. S.; Boatner, L. M.; Yen, S. L.; Cao, J.; Palafox, M. F.; Jami-Alahmadi, Y.; Backus, K. M. SP3-FAIMS Chemoproteomics for High-Coverage Profiling of the Human Cysteinome**. *ChemBioChem* **2021**, *22*, 1841–1851.
(41) Cao, J.; Boatner, L. M.; Desai, H. S.; Burton, N. R.; Armenta, E.; Chan, N. J.; Castellón, J. O.; Backus, K. M. Multiplexed CuAAC Suzuki–Miyaura Labeling for Tandem Activity-Based Chemoproteomic Profiling. *Anal. Chem.* **2021**, *93*, 2610–2618.
(42) Yang, F.; Jia, G.; Guo, J.; Liu, Y.; Wang, C. Quantitative Chemoproteomic Profiling with Data-Independent Acquisition-Based Mass Spectrometry. *J. Am. Chem. Soc.* **2022**, *144*, 901–911.
(43) Boatner, L. M.; Palafox, M. F.; Schweppe, D. K.; Backus, K. M. CysDB: A Human Cysteine Database Based on Experimental Quantitative Chemoproteomics. *Cell Chem. Biol.* **2023**, *30*, 683–698.e3.
(44) Clark, J. J.; Benson, M. L.; Smith, R. D.; Carlson, H. A. Inherent versus Induced Protein Flexibility: Comparisons within and between Apo and Holo Structures. *PLOS Comput. Biol.* **2019**, *15*, e1006705.
(45) Mariani, V.; Biasini, M.; Barbato, A.; Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **2013**, *29*, 2722–2728.
(46) Bryant, P.; Pozzati, G.; Elofsson, A. Improved Prediction of Protein-Protein Interactions Using AlphaFold2. *Nat. Commun.* **2022**, *13*, 1265.
(47) Hekkelman, M. L.; de Vries, I.; Joosten, R. P.; Perrakis, A. AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nat. Methods* **2023**, *20*, 205—-213.
(48) Lane, T. J. Protein structure prediction has reached the single-structure frontier. *Nat. Methods* **2023**, *20*, 170—173.
(49) Lazear, M. R.; Remsberg, J. R.; Jaeger, M. G.; Rothamel, K.; Her, H.-L.; DeMeester, K. E.; Njomen, E.; Hogg, S. J.; Rahman, J.; Whitby, L. R.; Won, S. J.; Schafroth, M. A.; Ogasawara, D.; Yokoyama, M.; Lindsey, G. L.; Li, H.; Germain, J.; Barbas, S.; Vaughan, J.; Hanigan, T. W.; Vartabedian, V. F.; Reinhardt, C. J.; Dix, M. M.; Koo, S. J.; Heo, I.; Teijaro, J. R.; Simon, G. M.; Ghosh, B.; Abdel-Wahab, O.; Ahn, K.; Saghatelian, A.; Melillo, B.; Schreiber, S. L.; Yeo, G. W.; Cravatt, B. F. Proteomic Discovery of Chemical Probes That Perturb Protein Complexes in Human Cells. *Mol. Cell* **2023**, *83*, 1725–1742.e12.
(50) Sen, S.; Young, J.; Berrisford, J. M.; Chen, M.; Conroy, M. J.; Dutta, S.; Di Costanzo, L.; Gao, G.; Ghosh, S.; Hudson, B. P.; Igarashi, R.; Kengaku, Y.; Liang, Y.; Peisach, E.; Persikova, I.; Mukhopadhyay, A.; Narayanan, B. C.; Sahni, G.; Sato, J.; Sekharan, M.; Shao, C.; Tan, L.; Zhuravleva, M. A. Small Molecule Annotation for the Protein Data Bank. *Database* **2014**, *2014*, bau116–bau116.
(51) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *PLoS Comput. Biol.* **2017**, *13*, e1005659.
(52) Lee, B.; Richards, F. M. The Interpretation of Protein Structures: Estimation of Static Accessibility. *J. Mol. Biol.* **1971**, *55*, 379–400.
(53) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinformatics* **2009**, *10*, 168.
(54) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423.
(55) de Brevern, A. G.; Bornot, A.; Craveur, P.; Etchebest, C.; Gelly, J.-C. PredyFlexy: Flexibility and Local Structure Prediction from Sequence. *Nucleic Acids Res.* **2012**, *40*, W317–W322.
(56) Daylight Theory: SMARTS - A Language for Describing Molecular Patterns. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.
(57) Chollet, F. Keras.
(58) Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A.; Žídek, A.; Green, T.; Tunyasuvunakool, K.; Petersen, S.; Jumper, J.; Clancy, E.; Green, R.; Vora, A.; Lutfi, M.; Figurnov, M.; Cowie, A.; Hobbs, N.; Kohli, P.; Kleywegt, G.; Birney, E.; Hassabis, D.; Velankar, S. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **2022**, *50*, D439–D444.