JACS Au
AN OPEN ACCESS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

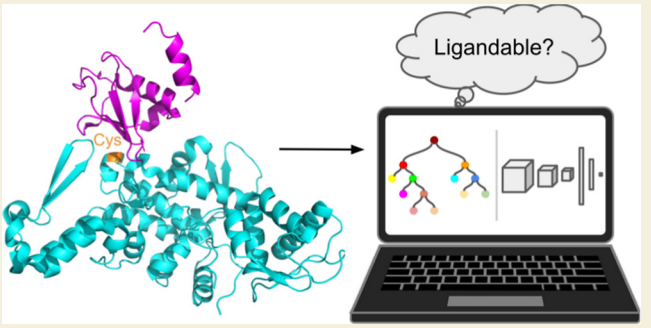
pubs.acs.org/jacsau
Article

Machine Learning Models to Interrogate Proteome-Wide Covalent Ligandabilities Directed at Cysteines

Ruibin Liu, Joseph Clayton, Mingzhe Shen, Shubham Bhatnagar, and Jana Shen*

Cite This: JACS Au 2024, 4, 1374−1384
Read Online

ABSTRACT: Machine learning (ML) identification of covalently ligandable sites may accelerate targeted covalent inhibitor design and help expand the druggable proteome space. Here, we report the rigorous development and validation of the tree-based models and convolutional neural networks (CNNs) trained on a newly curated database (LigCys3D) of over 1000 liganded cysteines in nearly 800 proteins represented by over 10,000 three-dimensional structures in the protein data bank. The unseen tests yielded 94 and 93% area under the receiver operating characteristic curves for the tree models and CNNs, respectively. Based on the AlphaFold2 predicted structures, the ML models recapitulated the newly liganded cysteines in the PDB with over 90% recall values. To assist the community of covalent drug discoveries, we report the predicted ligandable cysteines in 392 human kinases and their locations in the sequence-aligned kinase structure, including the PH and SH2 domains. Furthermore, we disseminate a searchable online database LigCys3D (https://ligcys.computchem.org/) and a web prediction server DeepCys (https://deepcys.computchem.org/), both of which will be continuously updated and improved by including newly published experimental data. The present work represents an important step toward the ML-led integration of big genome data and structure models to annotate the human proteome space for the next-generation covalent drug discoveries.



KEYWORDS: *covalent drug discovery, machine learning, protein structures, database, alphafold, kinases*

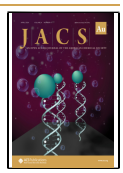
1. INTRODUCTION

Over the past two decades, targeted covalent inhibitor (TCI) discovery has become mainstream in the efforts to overcome 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 a traditional ligandable pocket for reversible binding.[4] An irreversible and sometimes also reversible covalent bond is formed between the warhead and a nucleophilic (reactive) amino acid residue in the target protein. Cysteine is the most nucleophilic amino acid and has been the most popular for covalent ligation. In fact, all FDA-approved TCIs are directed at a cysteine.[5]

In silico approaches hold great promise to accelerate proteome-wide TCI discovery efforts. In recent years, molecular dynamics (MD) simulations[6−8] have been developed to assess cysteine reactivities and ligandabilities; however, they cannot be scaled up to the proteome level due to the high computational cost. Machine learning (ML) models trained on the cysteine-liganded cocrystal structures in the protein data bank (PDB) have also been reported to evaluate the cysteine ligandabilities. In the earlier work, the support vector machine (SVM) models were trained on 1057 cysteine-liganded cocrystal structures of 515 proteins and achieved the best area under the curve of receiver operating characteristic (AUC the receiver operating characteristic or ROC), recall, and precision of 0.73, 0.62, and 0.41, respectively,[9] in an unseen test. In a most recent publication,[10] the graph neural network (GNN) models DeepCoSI were trained on the CovalentInDB database, which contains 10,042 cysteine-liganded cocrystal structures of 259 proteins, with the best training AUC of 0.92.

The emergence of the powerful and continuously improving AlphaFold2 (AF2) structure prediction engine[12] further underscores the potential utility of structure-based ligandability prediction tools. In this work, we present a new database LigCys3D (https://db.computchem.org/), which annotates 1133 ligandable cysteines of 778 proteins and their X-ray crystal structure representations in the PDB, including the

Received: November 28, 2023
Revised: February 22, 2024
Accepted: February 23, 2024
Published: April 5, 2024



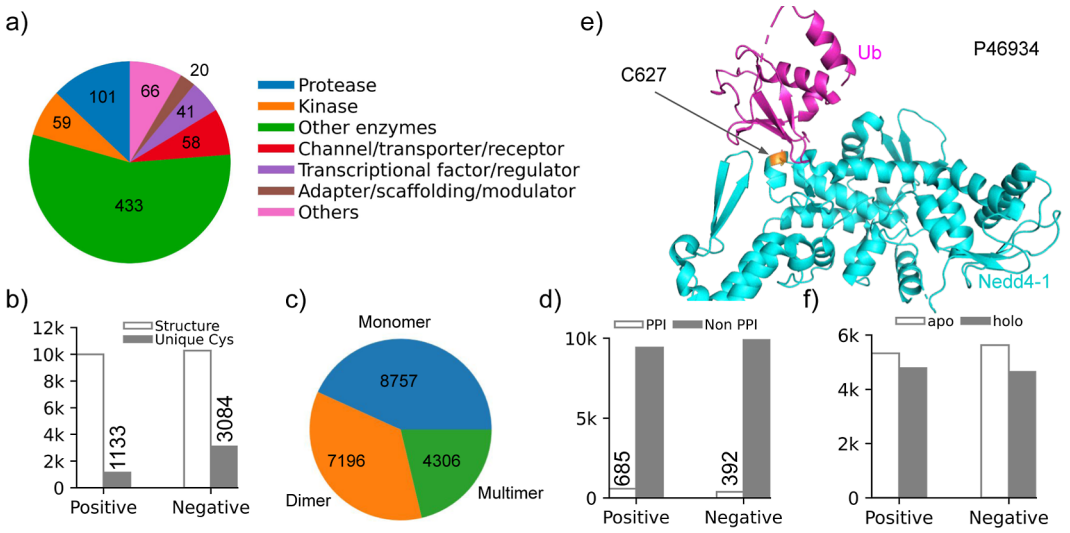


Figure 1. Analysis of the ligandable cysteines and the associated X-ray structures in the down-sampled LigCys3D. (a) Functional classes of the proteins that have at least one ligandable (positive) cysteine according to the structures deposited in the PDB. Functional information is taken from UniProtKB.[11] or SCOP2[15] (b) Number of positive and negative cysteines, and the number of PDB structures representing these cysteines. A positive cysteine is represented by up to 10 and an average of ∼9 structures, while a negative cysteine is represented by up to 4 and an average of ∼3 structures. (c) Number of nonunique cysteines that are in monomer, dimer, and multimer structures based on the biological assembly information in the PDB. (d) 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 sulfhydryl sulfur and the nearest heavy atom in another chain. (e) Nedd4-1 (cyan) contains a cysteine (C627, orange) at the PPI with ubiquitin (magenta) in the PDB entry5C7J. While not liganded in this structure, Cys627 is liganded by a covalent inhibitor in a different, monomeric structure (PDB ID: 5C91). (f) Number of PDB structures that are apo (ligand free) or holo (bound to any ligand) for positive and negative cysteines.

cysteine-liganded and unliganded forms. Employing this database, we trained and validated two types of ML models: the decision tree models and the three-dimensional convolutional neural networks (3D-CNNs). To our best knowledge, LigCys3D is the largest to date and significantly surpasses those used for the previous ML models[9,10] in terms of the number of unique cysteines, proteins, as well as the number of structural representations. Another unique feature of this work is the inclusion of ligand-free (apo) X-ray structures in the training set, which is expected to increase the extrapolation power of the ML models. Finally, the development of decision tree models (as opposed to the black box neural networks) based on physicochemical features allows us to dissect the physical determinants and gain systematic understanding of covalent ligandabilities. In multiple unseen tests, the XGBoost models and CNNs delivered the AUCs of 94% ± 1% and 93% ± 3%, respectively. The models were also applied to recapitulate the newly liganded cysteines based on AF2 predicted structures, giving recall values over 90%. Finally, a web server https://deepcys.computchem.org/ was developed to assist the community with covalent drug discovery.

2. RESULTS AND DISCUSSION

2.1. Construction of a Structure Database of Cysteine-Liganded Proteins Determined by Crystallography

In order to train the ML models, we first built a database of proteins containing cysteines that have been covalently modified by ligands. The recently published CovPDB[13] and CovalentInDB[10] databases together contain 659 liganded cysteines in 484 unique proteins. We performed an exhaustive search in the PDB and found an additional 474 liganded cysteines in 296 unique proteins. Together, we compiled 1133 liganded cysteines in 778 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 that 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,[14] we matched each cysteine with the (gene) accession number and residue ID in the UniProt knowledge base (UniProtKB).[11] 76% of the cysteine-liganded proteins are enzymes, including 101 proteases, 59 kinases, and 433 other enzymes (Figure 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[11] or SCOP2[15] (Figure 1a).

The CovPDB[13] and CovalentInDB[10] databases contain only the cysteine-liganded PDB structures, based on which the previously reported ML models were trained.[9,10] This is not ideal as the conformational variability is neglected, which may limit the model transferability (see a later discussion). Thus, we augmented the data set to a total of 10,105 positive entries (10,105 X-ray structures representing 1133 positive cysteines) and 97,754 negative entries (97,754 X-ray structures representing 3084 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 data set as LigCys3D (ligandable cysteine three-dimensional structure database) hereafter. Since there are significantly more negatives than positives, we randomly downsampled the negative entries to 10,267, i.e., 10,267 X-ray structures representing 3084 negative cysteines (an average of ∼3

https://doi.org/10.1021/jacsau.3c00749
JACS Au 2024, 4, 1374−1384

Table 1. Performance of the Tree-Based and CNN Models in the CVs and Unseen Tests[a]

metrics	ET		XGBoost		LightGBM		CNN	
	CV	test	CV	test	CV	test	CV	test
AUC	0.89 ± 0.00	0.94 ± 0.01	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	0.81 ± 0.01	0.90 ± 0.02	0.88 ± 0.01	**0.93±0.02**	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	0.74 ± 0.01	**0.91±0.01**	0.79 ± 0.01	0.94 ± 0.01	0.92 ± 0.02	**0.89±0.03**
selectivity	0.82 ± 0.01	0.82 ± 0.02	0.76 ± 0.02	0.77 ± 0.03	0.84 ± 0.01	0.85 ± 0.02	0.94 ± 0.02	0.69 ± 0.10
NPV	0.85 ± 0.00	0.75 ± 0.03	0.89 ± 0.01	0.81 ± 0.05	0.84 ± 0.00	0.70 ± 0.03	0.94 ± 0.01	0.86 ± 0.06
ACC	0.81 ± 0.00	0.88 ± 0.01	0.81 ± 0.01	0.89 ± 0.01	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 ± 0.01	**0.92±0.01**	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 CV or the unseen tests from 30 train, CV, and test experiments for ET and XGBoost, and 6 train, CV, and test experiments for LightGBM and CNN. Cysteines do not overlap between the training and test data sets. 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 are 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.

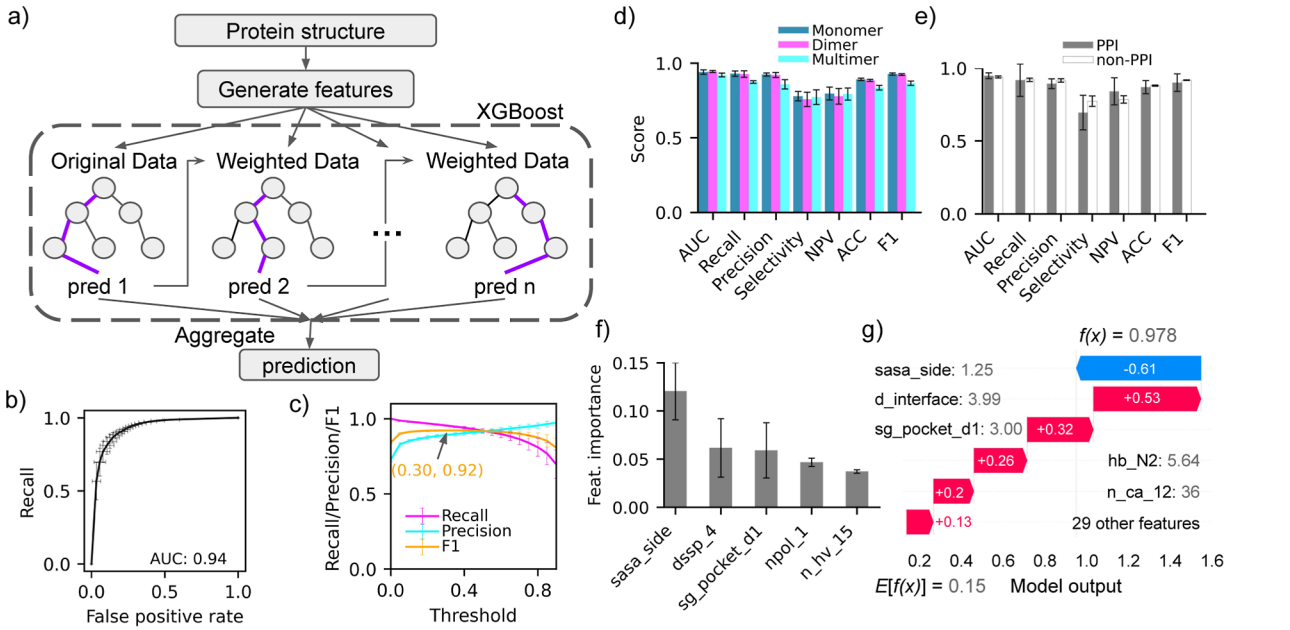


Figure 2. Performance of the tree-based models for predicting cysteine ligandabilities. (a) Model workflow based on the Extreme Gradient Boosting (XGBoost) classifier. (b) ROC curve for the XGBoost models obtained from 6 rounds of data splitting followed by training with 10-fold CV. The AUC is 0.94. (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 XGBoost models for cysteines in monomer, dimer, and multimer structures. (e) Performance metrics of the XGBoost models for cysteines at the interfaces (PPIs) or not. (f) Permutation-based feature importance scores for the top five features: the side chain cysteine SASA (sasa_side); secondary structure of the cysteine + 4 position (dssp_4); the distance from the cysteine sulfur to the nearest pocket (sg_pocket_d1); the minimum distance to any nonpolar atom in a different residue (npol_1); and the number of heavy atoms within 15 Å of the cysteine sulfur (n_hv_15). (g) "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.

structures per negative cysteine). In total, 20,259 entries were curated as the data set for model hold-out, training, and cross validation (CV, Figure 1b).

2.2. Structural Diversity, Variability, and Allostery Are Represented in the Augmented Data Set

Considering the quaternary structures associated with the entries, 8757 are monomers, 7196 are dimers, and 4306 are multimers (Figure 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, Figure 1d), as defined by using a distance cutoff of 4.5 Å between the cysteine sulfur and its nearest 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 (Figure 1e).[16] Nedd4-1 has a noncatalytic cysteine C627, which is located at the binding interface with ubiquitin (PDB ID: 5C7J)[17] and has been modified by a covalent inhibitor (PDB ID: 5C91).[16] In addition to the cysteine-liganded structures, through data augmentation, the positive cysteines are also represented by cocrystal structures in complex with reversible ligands as well as surprisingly more than 50% ligand-free structures. For the

https://doi.org/10.1021/jacsau.3c00749
JACS Au 2024, 4, 1374−1384

positive entries, 3912 are ligand-free and 3695 are ligand-bound, while for the negative entries, 3601 are ligand-free and 3003 are ligand-bound (Figure 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.[18] We should also note that in the LigCys3D data set, each protein has on average 1.5 ligandable cysteines, which suggests that allosteric sites are also represented.

2.3. 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.[6,8,19,20] Thus, we surmised that feature-based ML classification models such as decision trees may be suited for predicting cysteine ligandabilities. Based on the findings from these studies,[6,8,19,20] we devised a set of descriptors (37 after removal of multicollinearity, see the Materials and Methods section) for training the tree-based classifiers using PyCaret.[21] Given the small training data set, tree models avoid the overfitting problem that plagues the more sophisticated models that make use of vast parameter space. From the downsampled LigCys3D, 10% of the entries were randomly picked as holdouts for the "unseen" test, while the remaining 90% of the entries were reserved for training/CV. UniProt accession number and residue IDs were used to ensure that cysteines are unique between the training/CV and test sets. The 10-fold CV was used, where different folds have unique cysteines. Following CV, the model was retrained with hyperparameter tuning before being applied to the test set. This process (data splitting, training/CV, and test) was repeated six times to generate statistics for the model evaluation. To verify that the data splitting is unbiased and to generate more robust statistics, we also repeated the above process 30 times for the ET model. The resulting metrics are very similar, with the test AUC value unchanged (Table S1).

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 , 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 de-emphasized 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 XGBoost gave an AUC of 0.94 ± 0.01 (Figure 2b) and a maximum F1 score of 0.92 ± 0.02, which was achieved at the threshold value of 0.30 (Figure 2c). With this threshold, the recall and precision are 0.93 ± 0.02 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 the ligandable cysteines. Note, some test metrics of the tree models are higher than those of the CVs. This is because the CV metrics were calculated by averaging the models trained on 9-fold of data, while the test metrics were for the model trained on the entire 10-fold of data with the optimized hyperparameters.

2.4. 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 their proximity to interfaces (if any). We compared the XGBoost model performance metrics for cysteines in the monomer, dimer, and multimer structures (Figure 2d and Table S2). 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 is 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 (Figure 2e and Table S3). 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.[22]

2.5. Cysteine Ligandability Is Determined by a Set of Structural and Physicochemical 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 the cysteine ligandability. The feature importance score represents the decrease in the model score when a feature is randomly shuffled.[23] Accordingly, cysteine's side chain solvent-accessible surface area (sasa_side) is by far the most important feature (Figure 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.[24] An early bioinformatics analysis showed that cysteine is the least-exposed amino acid,[25] and the recent constant pH MD simulations showed that many hyperreactive cysteines in kinases[19,20] and other proteins[7] are buried. We will return 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 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,[26] which is supported by the fact that the front-pocket Ncap cysteine is the most popular site of targeted covalent inhibition among all kinases.[19] Similar to the BURIED term in the empirical pK_a prediction program PROPKA,[27] 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 SHapley Additive exPlanations (SHAP) values inform the impact of feature values on the prediction outcomes.[28,29] A positive or negative SHAP value increases

https://doi.org/10.1021/jacsau.3c00749
JACS Au 2024, 4, 1374−1384

or decreases the model output of a prediction from its expectation value estimated by randomly guessing from the features.[29] As an example, Figure 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.

2.6. Covalent Modification Perturbs the Cysteine Structure Environment

Structural perturbation by reversible ligands is a well-known phenomenon.[30] We hypothesized that covalent modification of a cysteine perturbs its structural environment. To test this hypothesis, we plotted the distributions of the cysteine side chain SASA and the distance to the nearest pocket, which are important features of the tree models (Figure 3). For the

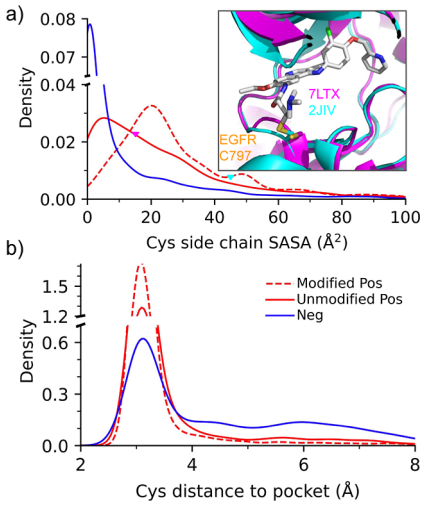


Figure 3. Cysteine's conformational environment is different between the modified and unmodified structures. (a) Distributions of the cysteine side chain SASA. Modified Pos (dashed red) and unmodified Pos (solid red) refer to the positive structures in which the cysteine is liganded and unliganded, respectively. The ligand was removed in the SASA calculations. The inset shows an overlay between the unmodified (PDB: 7LTX) and modified (PDB: 2JIV) X-ray structures of the EGFR kinase. The front-pocket C797 has a SASA value of 16.9 and 44.6 Å^2 in the apo and holo states, respectively. (b) Distributions of the distance from the cysteine sulfur to the nearest pocket (alpha sphere).

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. Interestingly, 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 close to the peak of the negatives (near zero) (Figure 3a). This analysis suggests that covalent modification perturbs the protein structure so as to increase cysteine's solvent exposure. Furthermore, while a larger fraction of ligandable cysteines are solvent exposed as compared to the unligandable cysteines, a significant fraction of ligandable cysteines are actually deeply buried. The latter is consistent with the notion that cysteine is the least solvent-exposed amino acid[25] and our recent finding that most reactive cysteines in kinases[8,19,20] and other proteins are in fact buried.[7]

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, Figure 3b); however, the modified positives have a higher peak intensity, suggesting that covalent ligand binding may slightly "pull" the cysteine toward 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.

2.7. Importance of Including Cysteine-Unmodified Structures in the Training Set

Given that covalent ligation perturbs the structure such that the difference in the cysteine environment between the modified positives and negatives is larger than that between the unmodified positives and negatives (Figure 3), we hypothesized that models trained using the modified structures may give a "deceptively" higher performance than models trained with the unmodified structures. To test this, we compare the performances of the ET models trained with the modified (model 1), unmodified (model 2), and combined (model 3) structures. Confirming our hypothesis, the unseen test AUC, recall, and precision of model 1 are all above 0.95, whereas the AUC of model 2 dropped to 0.85 and recall/precision to 0.75/0.78 (Tables 2 and S6). Even though the training data set of model 3 is the largest, 68% larger than model 1, the AUC is slightly lower at 0.94 and recall/precision is 0.89/0.93.

Table 2. Impact of Training with Unmodified Structures on the ET Model Predictions[a]

model[b]	model 1	model 2	model 3
structures	modified	unmodified	combined
Pos/Neg[c]	5931:5931	4061:4061	9992:10,267
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

[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 data set (training, CV, and unseen test).

We compare the prediction metrics of model 1 with the published metrics of the previous two models[9,10] that were trained using a smaller data set and the cysteine-liganded structures only. Note, this comparison is for information only and needs to be taken with a grain of salt due to the difference in the training and test data. The ET metrics (AUC, recall, and precision all above 0.95) surpass the feature-based SVM model (test AUC, recall, and precision of 0.73, 0.62, and 0.41),[9] which may be attributed to the larger data set and the use of

https://doi.org/10.1021/jacsau.3c00749
JACS Au 2024, 4, 1374−1384

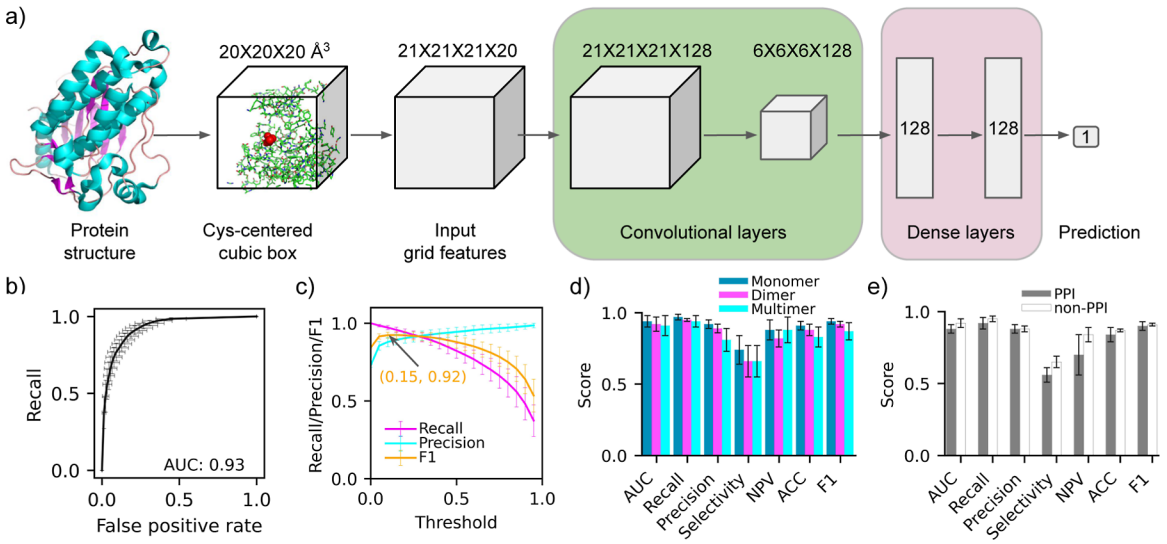


Figure 4. Performance of the three-dimensional CNN. (a) Architecture of the 3D-CNN for cysteine ligandability predictions. (b) ROC curve obtained from 6 train/test experiments. The AUC is indicated for the average curve. (c) Recall/precision/F1 score as a function of the classification threshold. The 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.

the physio-chemical features born out of our previous mechanistic studies of cysteine reactivities and ligandabilities.[6,19,20] The ET model's test AUC (0.96) is also slightly higher than the training validation AUC (0.92) of the most recent GNN model (other metrics were not reported).[10]

Since in prospective predictions modified structures are unavailable, we asked if the models should be trained with the unmodified structures only. To address this question, we applied the three ET models to the ligandable cysteines discovered in an early chemoproteomics experiment conducted in cell lysates.[31] These cysteines are not in LigCys3D; i.e., they do not have modified structures. As expected, model 1 has by far the lowest recall; however, model 3 is slightly better than model 2 in both recall and precision (Tables 2 and S6). This analysis demonstrates that structure variation (including both modified and unmodified structures) further enhances the extrapolation power of the models.

2.8. CNN Models Show a 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[32] and recently adapted for protein pK_a predictions.[33] In our modified architecture, a cubic grid of 20 × 20 × 20 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 (Figure 4a). 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 (Figure 4a, details see the Materials and Methods section). To allow comparison to the tree models, data splitting and CV were conducted in the same manner. Interestingly, the 3D-CNN gave very similar results to the best tree model XGBoost, with the AUC, accuracy, and precision of 0.93 ± 0.04, 0.96 ± 0.02, and 0.89 ± 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 (Figure 4b). 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, Figure 4c).

We also examined the CNN performance for different protein quaternary structures and PPI vs non-PPI cysteines in comparison to the XGBoost models (Figure 4d and Table S4). 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 those between the PPI and non-PPI cysteines (Figure 4e and Table S5). 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 model prediction. In contrast, the features used in the tree models cover all residues in the bioassembly regardless the distances to the cysteine of interest.

https://doi.org/10.1021/jacsau.3c00749
JACS Au 2024, 4, 1374−1384

2.9. External Evaluation on the Newly Published Ligandable Cysteines Captured by X-ray Crystallography

To further evaluate the ET and CNN models, they were tested on a new, nonoverlapping data set of 30 unique proteins with 38 covalently liganded cysteines, for which the cocrystal structures were deposited in the PDB after the LigCys3D was constructed in Nov 2022. To demonstrate the general utility of the ML models, the AF2 structures were used for making predictions. The ET model recapitulated 35 out of 38 (a recall of 0.92), while the CNN recapitulated 34 out of 38 liganded cysteines (a recall of 0.90, see Table S7). Curiously, both the ET and CNN missed C269 in isocitrate dehydrogenases 1 (IDH1, UniProt ID O75874). Closer examination revealed that the covalently liganded IDH1 is an R132H mutant, whereas the AF2 structure used for the predictions is of the wild type (WT). Comparison of the deposited X-ray structure of the R132H mutant (PDB ID: 8HB9)[34] with the WT AF2 structure showed that the pocket near C269 is enlarged in the mutant. This difference could explain the false negative predictions. Another cysteine that both ET and CNN models missed is C412 of the YTHDF1 protein (UniProt ID Q9BYJ9). In the deposited X-ray structure (PDB ID: 7PCU).[35] C412 is fully exposed to solvent; however, in the AF2 structure, it is fully buried as a nearby loop (residues 95−114) is collapsed onto the pocket where the cysteine resides. These two cases show that the accuracy of the ligandability prediction is dependent on the quality of the AF2 predicted structure, which is a limitation of structure-based ML models. Effects of mutations and loop modeling are challenging for the current AF2 model, and the ligandability predictions will benefit from the continued improvement of the AlphaFold structure prediction engine.

2.10. Prospective Prediction of Ligandable Cysteines in the Human Kinome

Human kinases are important drug targets, and most FDA-approved covalent drugs are kinase inhibitors directed at a cysteine in the catalytic or allosteric pocket of the kinase domain. LigCys3D contains 46 unique liganded cysteines in 37 kinases, and all but two are in the kinase domain (red circles in Figure 5). The front pocket αD helix, with 10 unique cysteines belonging to different kinases, is the most targeted location. This is consistent with the constant pH MD simulations, which showed that the cysteine at or near the N-terminal cap of the αD helix is hyper-reactive due to the local hydrogen bonding and electrostatic environment.[19] Other popular locations for cysteine ligation are the N-terminal part of the activation loop (ALN, 8 cysteines), αE (5 cysteines), and the C-terminal end of βI (βIC or p-loop, 4 cysteines).

The human kinome contains 536 kinases;[37] however, only a small fraction (37 according to the cocrystal structures in the PDB) have been covalently drugged. Thus, we applied the ML models based on the AF2 structures to make predictions of the ligandable cysteines in the rest of the 481 kinases, accounting for not only the kinase domain but also the PH, SH1, and SH2 domains. Note, 18 kinases that do not have AF2 structure models in the EBI repository were excluded, and since no ligandable cysteines were predicted in the SH1 domain, it is excluded in the discussion below.

A total of 1083 cysteines in the kinase, PH, and SH2 domains of 392 kinases were predicted to be ligandable, and 89 kinases were predicted to have no ligandable cysteine. Figure 5 shows the number of the predicted ligandable cysteines and their locations in a kinase structure according to the multiple-sequence alignment of Modi and Dunbrack.[36] Note, 79 predicted ligandable cysteines in 28 kinases are excluded in Figure 5 due to the lack of sequence alignment information. Three kinase locations with the most positive cysteines are ALN, the C-terminal end of the activation loop (ALC), and αH (Figure 5). Other interesting locations include αD, αE, βIC, and αC. All of these locations have been covalently targeted in other kinases (according to LigCys3D). Thus, our kinome-wide predictions demonstrate new opportunities for covalent kinase inhibitor design but also suggest potential selectivity problems. Interestingly, the ML predictions uncovered 7 previously untargeted locations that contain predicted ligandable cysteines, including the PH domain, the N-terminal end of βI (βIN), βIII−αC, βIV−βV, αE−βVI, FL-αG, and αG−αH (Figure 5). Many of these locations are in the loops connecting two secondary structures.

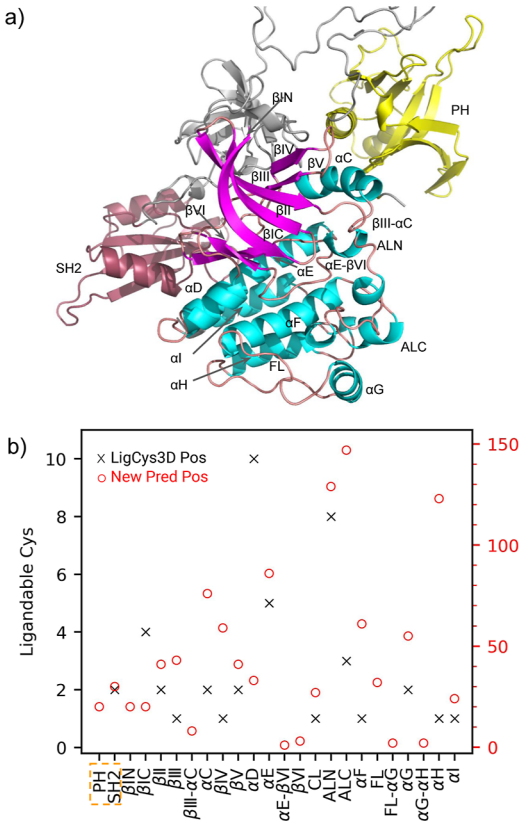


Figure 5. Ligandable cysteines in the kinase, PH, and SH2 domains of the human kinases. (a) The PH (yellow), SH2 (dark red), and catalytic (magenta for β strands and cyan for helices) domains of a kinase are shown in a representative AF2 model structure of BTK (Uniprot ID: Q06187). The loops and regions not in the PH, SH2, or kinase domains are colored gray. The various structure elements in the kinase domain (named according to Modi and Dunbrack[36]) as well as the PH and SH2 domains are labeled. (b) The number of liganded cysteines in LigCys3D (gray crosses, left y axis) and the predicted ligandable cysteines for the 397 (typical) human kinases not in LigCys3D (red circles, right y axis) mapped onto the various structure elements of the kinase domain as well as the PH and SH2 domains (dashed box). The multiple sequence alignment based on the kinase domain is taken from Modi and Dunbrack.[36] The consensus predictions by the ET, XGBoost, and LightGBM models are shown. The domain and structure element names are ordered by the sequence and only those with at least one ligandable cysteine are shown (no ligandable cysteines were predicted for the SH1 domain).

https://doi.org/10.1021/jacsau.3c00749
JACS Au 2024, 4, 1374−1384

2.11. Development of a Live Online Database (LigCys3D) and Prediction Server (DeepCys)

To assist the community of TCI discoveries, we implemented a searchable web database LigCys3D (https://ligcys.computchem.org/). Each row or entry is a ligandable cysteine in a protein, along with the following information as columns: the PDB ID, chain ID, PDB residue ID, Uniprot ID, Uniprot residue ID, bioassembly type, the calculated side chain SASA, distance to a protein−protein interface, distance to the nearest pocket, as well as whether the cysteine is covalently modified by a ligand in the particular structure (PDB file). Any information specified in the column can be filtered, and a list of matched entries are generated. The user can download the matched entries or the entire database as a CSV file. The database will be continuously updated to include newly discovered ligandable cysteines in the PDB.

We also implemented a web server DeepCys (https://deepcys.computchem.org/) that can predict cysteine ligandability given a PDB ID, Uniprot ID, or a structure file in the PDB format. Provided a PDB ID, the web server queries the RCSB repository for the bioassembly file. Provided a Uniprot ID, the web server queries the AlphaFold Protein Structure Database of EMBL-EBI (https://alphafold.ebi.ac.uk/) for a corresponding AF2 predicted structure. Once a structure file is retrieved or provided, the server uses the models trained in this study (which will be continuously improved based on the continuously expanding training data) to make ligandability predictions for all cysteines in the protein. The results are in the format of a CSV file, which contains the chain ID, residue ID, classification probability, and ligandability prediction for each cysteine. The CSV file will be sent to the user-provided email address.

3. CONCLUSIONS

Exploiting a newly curated database (LigCys3D) of ∼1000 liganded cysteines in ∼800 proteins represented by ∼10,000 three-dimensional structures in the PDB, we developed the tree-based and 3D-CNN models for proteome-wide cysteine ligandability predictions. In multiple unseen tests, the ET and XGBoost models gave an AUC of 94%, while the CNN models gave an AUC of 93%. Based on the AF2 predicted structures, the ET model and CNN recapitulated the newly liganded cysteines with recall values of 92 and 90%, respectively. The fact that the tree models have orders of magnitude smaller parameter space (i.e., less than 40 structural and physicochemical features) than CNNs reinforces the notion that the reactivity and ligandability of cysteines are largely determined by the structure environment, solvent accessibility, and potential hydrogen bonding as well as electrostatic interactions. Also, the tree models avoid the overfitting problem facing the more sophisticated ML models that make use of very large parameter space.

Despite the promising test results, the models have several limitations that need to be addressed in future work. First, the training data set is very small and underrepresents transmembrane proteins, transcription factors, and other nonenzymes. Second, the potential effects of mutation and posttranslational modification on the structure as well as the membrane environment of transmembrane proteins are not accounted for. Third, the current AF2 engine offers a very limited accuracy for predicting loop conformations. Finally, the model performance can perhaps be more accurately assessed by using sequence-based data splitting. Nonetheless, the present work represents an important step toward the ML-led integration of big genome and structure data to annotate the human proteome space for the next-generation covalent drug discoveries. To assist the community in the covalent drug design efforts, we report the predicted ligandable cysteines in 392 human kinases and their locations in the sequence-aligned kinase structure, including the PH and SH2 domains. Furthermore, we disseminate a web database (https://ligcys.computchem.org/) and a web prediction server (https://deepcys.computchem.org/), both of which will be continuously updated and improved by including newly published experimental data.

4. MATERIALS AND METHODS

4.1. Construction of the LigCys3D Database

Two recently published databases, CovPDB[13] and CovalentInDB,[10] compiled cysteine-liganded cocrystal structures in the RCSB Protein Data bank (PDB). These two databases have overlap, and together they provide 2875 cysteine-liganded cocrystal 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"[38] 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[14] to annotate the liganded cysteines with UniProt accession numbers and residue IDs (https://www.uniprot.org),[11] 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).[39] We refer to this data set as LigCys3D.

4.2. Data Engineering for the ML Models

To construct an ML training set with balanced positive and negative classes and to reduce model training time, we downsampled the number of structures in LigCys3D as follows. For each positive cysteine (based on the UniProt accession number and residue ID), all cysteine-liganded structures were included, and the cysteine-unliganded structures were selected using a SASA-based protocol (see below) such that the total number of structures does not exceed 10. The cysteine-liganded and unliganded structures are termed the liganded and unliganded positives, respectively. For each negative cysteine, all and up to a total number of 4 structures were selected using a similar SASA-based protocol (see below). To maximize structural variation, the unliganded positive structures were put into four bins based on the cysteine side chain 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 training data set (downsampled from LigCys3D) was constructed, comprising 9992 positive (1133 unique positive cysteines in 9992 structures) and 10,267 negative (3084 unique negative cysteines in 10,267 structures) entries. We will use this data set for model training and testing.

4.3. Feature Engineering for the Tree Models

Features are critical for the performance of tree-based models. We conceived a set of structural and physicochemical features based on our findings from the constant pH MD analysis of cysteine reactivities and ligandabilities in a large number of kinases[6,8,19,20] and other enzymes.[7] In total, eights types of features were calculated based on the input structure, including solvent accessibility (proximity to

https://doi.org/10.1021/jacsau.3c00749
JACS Au 2024, 4, 1374−1384

hydrophobic residues and the cysteine SASA calculated with NACCESS[40]); distance to pockets (defined by fpocket[41]); potential hydrogen bonding; electrostatic interactions; secondary structures (calculated with Biopython[42]); residue flexibility (calculated with PredyFlexy[43]); 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 Supporting Information Methods. After removal of highly correlated features, 37 features were left (see Supporting Information Methods).

4.4. Data Splitting and Training of the Tree Models

PyCaret[21] was used for building tree-based classifiers. We manually separated 10% of the data as the unseen test set and 90% as the training set (see below). The training set was used for 10-fold 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 CV, the groupKfold method was used to avoid placing (modified or unmodified) structures associated with the same cysteine in different folds.

Although ligandability of a cysteine is mainly determined by its local conformational environment,[6−8] cysteines in two highly similar structures resulting from two similar sequences may have similar ligandabilities. To accurately calculate the sequence similarity between two proteins, sequence alignment needs to be performed, which would be a prohibitively large effort for our data set. In the previous ML model development,[9,10] sequence-based clustering of proteins by cd-hit[44] was used in data splitting; however, the cd-hit method[44] is based on simple word (small stretch of sequence) counting and not an actual sequence alignment. Thus, we opted to skip the sequence clustering step and rely on multiple random data splittings in model evaluation. This is a limitation, which should be addressed in future work.

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 a 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.

4.5. Training of CNNs

The test-train splitting and 10-fold CV were performed in the same manner as for the tree models. The 3D-CNN architecture was adapted and modified from the Pafnucy model,[32] which was recently adapted for protein pK_a predictions.[33] 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[45] 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 the rotational variance.

Keras 2[46] 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 are flattened by a 128 units Dense layer with relu activation, normalized by a BatchNormalization 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 are 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 new AF2 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. For the unseen and external testing, the predictions were determined by majority voting.

4.6. External Validation on the Newly Liganded Cysteines

Structure files that were deposited in the PDB between 11/29/2022 and 10/11/2023 and have publications were retrieved and analyzed as to (1) whether the cysteines are covalently modified as defined above; (2) whether the corresponding protein is present in LigCys3D; and (3) whether an AF2 structure corresponding to the Uniprot accession number is available in the EMBL-EBI AlphaFold repository (https://alphafold.ebi.ac.uk/). The AF2 structures were subsequently used in the ligandability predictions.

4.7. Prediction of Ligandable Cysteines in the Human Kinome

We collected the human kinase list from KinMap.[37] For those with AF2 predicted structures, ligandability predictions were made by using three tree methods (ET, XGBoost, and LightGBM) and the consensus scheme. The kinase domain information was extracted from the Uniprot Family & Domains section and if the domain name starts with text Kinase domain, the corresponding residue range is considered as a kinase domain for that Uniprot ID. We also used the KinCoRe alignment file[36] downloaded from the website http://dunbrack.fccc.edu/kincore/biojs to assign the structure location for each cysteine in the kinase domain. For cysteines in the PH, SH2, and SH1 domains, if the Uniprot domain name starts with them, then the original domain names are renamed as the short forms so that multiple SH2 domains are combined. For those cysteines that are in the kinase domain but are not found in KinCore, the domain name is just Kinase domain but they are not shown in Figure 5. Cysteines not in the kinase, PH, SH1, and SH2 domains are not discussed in this study. For the location names in KinCore, A was replaced with α to indicate α-helix, B was replaced with β to indicated β-sheet, and an Arabic number was replaced by a Roman number. These changes were made to be more consistent with those in the literature.

4.8. Calculation of Model Performance Metrics

Given a confusion matrix composed 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.

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

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

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

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

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

$$\text{F1} = 2 \times \text{recall} \times \text{precision}/(\text{recall} + \text{precision}) \quad (6)$$

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

https://doi.org/10.1021/jacsau.3c00749
JACS Au 2024, 4, 1374−1384

■ ASSOCIATED CONTENT

Data Availability Statement

All training and testing data as well as the ML models are downloadable from https://github.com/JanaShenLab/DeepCys. The training data set LigCys3D is also provided on the web as a searchable database (https://ligcys.computchem.org/). The tree models and CNNs are also provided as a web server DeepCys (https://deepcys.computchem.org/).

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacsau.3c00749.

Additional methods, tables, and figures (PDF)

Predicted ligandable cysteines in 392 kinases and their locations provided as a downloadable file (Kinase_Ligandable_Cys.xlsx) (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Jana Shen − *Department of Pharmaceutical Sciences, University of Maryland School of Pharmacy, Baltimore, Maryland 21201, United States;* orcid.org/0000-0002-3234-0769; Email: jana.shen@rx.umaryland.edu

Authors

Ruibin Liu − *Department of Pharmaceutical Sciences, University of Maryland School of Pharmacy, Baltimore, Maryland 21201, United States*

Joseph Clayton − *Department of Pharmaceutical Sciences, University of Maryland School of Pharmacy, Baltimore, Maryland 21201, United States; Division of Applied Regulatory Science, Office of Clinical Pharmacology, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, Maryland 20993, United States;* orcid.org/0000-0002-2652-5994

Mingzhe Shen − *Department of Pharmaceutical Sciences, University of Maryland School of Pharmacy, Baltimore, Maryland 21201, United States;* orcid.org/0000-0001-7461-3764

Shubham Bhatnagar − *Department of Computer Science, University of Maryland at College Park, College Park, Maryland 20742, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/jacsau.3c00749

Author Contributions

CRediT: **Ruibin Liu** data curation, formal analysis, investigation, methodology, software, validation, visualization, writing-original draft, writing-review & editing; **Joseph Clayton** formal analysis, investigation, methodology, writing-review & editing; **Mingzhe Shen** data curation, formal analysis, investigation; **Shubham Bhatnagar** software; **Jana Shen** conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, resources, supervision, validation, visualization, writing-original draft, writing-review & editing.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

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 Discovery 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. K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions. *Nature* **2013**, *503*, 548−551.

(5) Singh, J. The Ascension of Targeted Covalent Inhibitors. *J. Med. Chem.* **2022**, *65*, 5886−5901.

(6) 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.

(7) 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.

(8) 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.

(9) Zhang, W.; Pei, J.; Lai, L. Statistical Analysis and Prediction of Covalent Ligand Targeted Cysteine Residues. *J. Chem. Inf. Model.* **2017**, *57*, 1453−1460.

(10) 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.

(11) 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.; Masson, P.; et al. UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480−D489.

(12) 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.

(13) 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.
(14) 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.
(15) 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.
(16) 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.
(17) 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.
(18) 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.
(19) 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.
(20) 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.
(21) Ali, M. *PyCaret: An Open Source, Low-Code Machine Learning Library in Python*, 2020.
(22) 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. *Signal Transduction Targeted Ther.* **2020**, *5*, 213.
(23) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.
(24) 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. Immunol. Methods* **2008**, *332*, 41−52.
(25) Marino, S. M.; Gladyshev, V. N. Cysteine Function Governs Its Conservation and Degeneration and Restricts Its Utilization on Protein Surfaces. *J. Mol. Biol.* **2010**, *404*, 902−916.
(26) 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.
(27) 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.
(28) Shapley, L. S. A. *Value for N-Person Games*; RAND Corporation, 1952.
(29) Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2017.
(30) 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*, No. e1006705.
(31) 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.
(32) 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.
(33) Cai, Z.; Luo, F.; Wang, Y.; Li, E.; Huang, Y. Protein p K_a Prediction with Machine Learning. *ACS Omega* **2021**, *6*, 34823−34831.
(34) Liang, Q.; Wang, B.; Zou, F.; Guo, G.; Wang, W.; Wang, W.; Liu, Q.; Shen, L.; Hu, C.; Wang, W.; Wang, A.; Huang, T.; He, Y.; Xia, R.; Ge, J.; Liu, J.; Liu, Q. Structure-based discovery of IHMT-IDH1−053 as a potent irreversible IDH1 mutant selective inhibitor. *Eur. J. Med. Chem.* **2023**, *256*, 115411.
(35) Micaelli, M.; Dalle Vedove, A.; Cerofolini, L.; Vigna, J.; Sighel, D.; Zaccara, S.; Bonomo, I.; Poulentzas, G.; Rosatti, E. F.; Cazzanelli, G.; Alunno, L.; Belli, R.; Peroni, D.; Dassi, E.; Murakami, S.; Jaffrey, S. R.; Fragai, M.; Mancini, I.; Lolli, G.; Quattrone, A.; Provenzani, A. Small-Molecule Ebselen Binds to YTHDF Proteins Interfering with the Recognition of N 6-Methyladenosine-Modified RNAs. *ACS Pharmacol. Transl. Sci.* **2022**, *5*, 872−891.
(36) Modi, V.; Dunbrack, R. L. A Structurally-Validated Multiple Sequence Alignment of 497 Human Protein Kinase Domains. *Sci. Rep.* **2019**, *9*, 19790.
(37) Eid, S.; Turk, S.; Volkamer, A.; Rippmann, F.; Fulle, S. KinMap: a web-based tool for interactive navigation through human kinome data. *BMC Bioinf.* **2017**, *18*, 16.
(38) 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.
(39) 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*, No. e1005659.
(40) Lee, B.; Richards, F. M. The Interpretation of Protein Structures: Estimation of Static Accessibility. *J. Mol. Biol.* **1971**, *55*, 379.
(41) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinf.* **2009**, *10*, 168.
(42) 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.
(43) 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.
(44) Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data. *Bioinformatics* **2012**, *28*, 3150−3152.
(45) *Daylight Theory: SMARTS—A Language for Describing Molecular Patterns*. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.
(46) Chollet, F. Keras. https://keras.io/2.15/api/, (accessed January 2024).

https://doi.org/10.1021/jacsau.3c00749
JACS Au 2024, 4, 1374−1384