CovCysPredictor: Predicting Selective Covalently-Modifiable Cysteines using Protein Structure and Interpretable Machine Learning

Bryn Marie Reimer[1,2*], Ernest Awoonor-Williams[1], Andrei A. Golosov[1], Viktor Hornak[1*]

[1]Novartis Biomedical Research, Computer-Aided Drug Discovery, Global Discovery Chemistry, 181 Massachusetts Avenue, MA 02139, United States.

[2]University of Massachusetts Amherst, Manning College of Information & Computer Sciences, 140 Governors Drive, Amherst, MA 01003, United States.

* Co-corresponding authors

Cysteine, Covalent Modification, Machine Learning, Proteome, Covalent Drug Discovery

ABSTRACT: Targeted covalent inhibition is a powerful therapeutic modality in the drug discoverer's toolbox. Recent advances in covalent drug discovery, in particular targeting cysteines, have led to significant breakthroughs for traditionally challenging targets such as mutant KRAS, which is implicated in diverse human cancers. However, identifying cysteines for targeted covalent inhibition is a difficult task, as experimental and *in silico* tools have shown limited accuracy. Using the recently released CovPDB and CovBinderInPDB databases, we have trained and tested interpretable machine learning models to identify cysteines which are liable to be covalently modified (*i.e.,* "ligandable" cysteines). We explored myriad physicochemical features (pK_a, solvent exposure, residue electrostatics, *etc*.) and protein-ligand pocket descriptors in our machine learning models. Our final logistic regression model achieved a median F_1 score of 0.73 on held-out test sets. When tested on a small sample of *holo* proteins, our model also showed reasonable performance, accurately predicting the most ligandable cysteine in most cases. Taken together, these results indicate that we can accurately predict potential ligandable cysteines for targeted covalent drug discovery, privileging cysteines which are more likely to be selective rather than purely reactive. We release this tool to the scientific community as CovCysPredictor.

INTRODUCTION

Covalent modification of nucleophilic residues in proteins is gaining traction as a high-value strategy in drug discovery.[1–9] Among the nucleophilic amino acid residues in proteins, cysteine is the most nucleophilic, with its thiol sidechain playing diverse functional roles in biochemistry.[10,11] The nucleophilicity of cysteines has been exploited in designing covalent drugs,[12–19] which tend to be therapeutically more potent with longer residence time than their non-covalent counterparts. Additionally, cysteine's nucleophilicity presents opportunities for tackling traditionally 'intractable' targets via novel chemical modalities, such as targeted protein degradation.[20] However, not all cysteines in proteins are suitable for covalent modification. For instance, in Bruton's Tyrosine Kinase (BTK) — a well-validated clinical target for the treatment of multiple B-cell cancers — there are three cysteines (C462, C481, and C527), but only C481 has been successfully targeted by several FDA-approved small molecule drugs, such as ibrutinib (**Figure 1**). Thus, identifying cysteine residues susceptible to covalent modification could inform targeted approaches in covalent drug discovery.

Predicting the reactivity of a cysteine residue or its propensity towards covalent modification is a challenging task,[21] which depends on a number of complex factors including pK_a, accessible surface area, acidity, residue microenvironment, steric and electrostatic effects, and hydrogen bonding interactions.[22,23] A given cysteine's pK_a is widely variable depending on the neighboring residue microenvironment,[21,24] with experimentally measured cysteine pK_a values ranging from 2.9 to 11.1.[25] A measure of the reactivity of a Cys residue can be deduced from its pK_a,[26] which informs the relative proportion of charged thiolate to neutral thiol species. Cysteines with a lower pK_a have a greater proportion of reactive thiolate species available for chemical protein modification. Traditional structure-based *in silico* methods are often limited in their predictive performance of experimental cysteine pK_a,[24,27] and accurate prediction of cysteine pK_a residues in proteins remains an unsolved problem in computational biochemistry.[28] Moreover, there are limited comprehensive datasets of experimental cysteine pK_a values, making rigorous validation and comparison of these methods difficult.

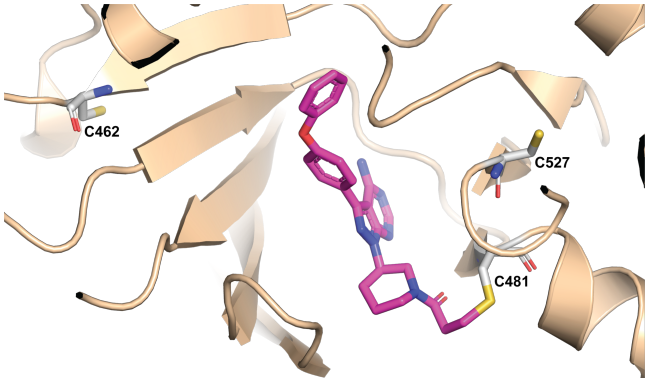


Figure 1. Structure of BTK in complex with anticancer drug, ibrutinib, highlighting the different cysteine locations in the protein (PDB ID: 5P9J). Ibrutinib covalently binds to C481 of BTK.

Despite these challenges, attempts have been made to predict cysteine pK_a values. Most *in silico* approaches have relied on a combination of physicochemical features (e.g., residue pK_a, solvent exposure, *etc*.) from a structure-based perspective,[29–38] and sometimes sequence-based information[39] to assess the reactivity or ligandability of cysteine residues in protein targets.[40] Chemoproteomic approaches such as quantitative reactivity profiling aimed at labeling reactive cysteines at a proteome-wide scale have also been developed;[15,41] however, identifying site-selective cysteines appropriate for targeted covalent drug discovery efforts from such approaches is not straightforward.[42] Earlier work by Soylu and Marino[29] saw the development of an algorithm and webserver to predict cysteine reactivity. Their algorithm combined energy-based hydrogen-bond network contributions with knowledge-based sequence profiling approaches to access cysteine reactivity in proteins.[29] Other studies have performed statistical analysis on covalent modified cysteines to identify features that allow ligandable cysteines to be easily detected.[31] More recent approaches based on deep learning have aimed to abstract the problem to a graph-based neural network, which does not require explicit pK_a or solvent exposure calculations.[35] However, the model is sensitive to small perturbations in the positions of sidechains and, like many deep learning models, has parameters that are difficult to interpret.

In this work, we utilized new datasets available for structures of covalently-modifiable residues (COVPDB[43] and CovBinderInPDB[44]), focusing on cysteine residues. We assessed computational methods for protein structure preparation and residue property prediction to improve upon the classical models of cysteine reactivity which take residue pK_a and solvent exposure into account. Based on this information and with the inclusion of additional features such as local amino acid environment and protein binding pocket information, we built a machine learning model to predict potential covalently-modifiable cysteines (i.e., "ligandable" cysteines) from protein structural data. Our highly interpretable logistic regression model shows good performance, comparable to other "black-box" predictive machine learning models in the literature.[31,35] Our results suggest that we can accurately predict ligandable cysteines from protein structure information, which is especially useful in accelerating the pace of early drug discovery efforts towards targeted covalent inhibitor design for clinical therapeutic benefit.

MATERIALS AND METHODS

Sources of Structural Data

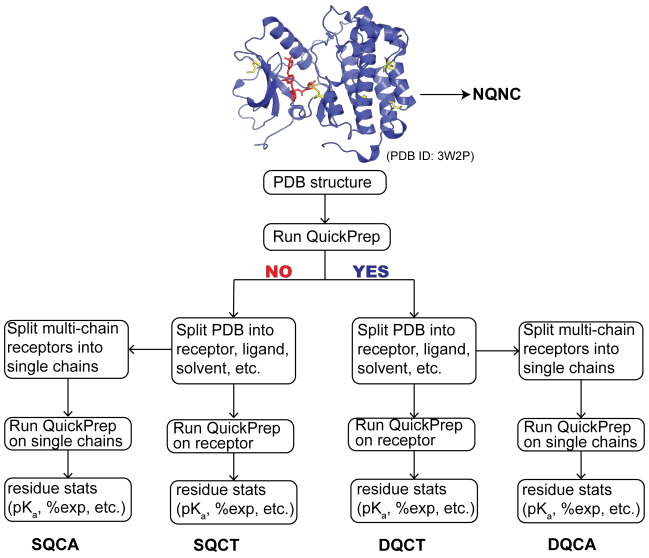
The structural data for proteins containing covalently modified residues were downloaded from the COVPDB[43] and CovBinderInPDB[44] databases. The COVPDB database consists of 959 high-resolution 3-dimensional biologically relevant covalent protein-ligand complexes with ligands bound to cysteine, mined from the Protein Data Bank (PDB).[45] In total, 291 unique protein chains are represented; 748 ligands; and 59 warheads. The CovBinderInPDB database is a more recent structure-based covalent binder database which contains 1344 complexes containing a covalently modified cysteine. CovBinderInPDB has 421 unique protein chains, 1009 covalent binders, and 66 warheads represented in its database.[44] We compiled datasets derived from both COVPDB and from CovBinderInPDB. Cysteine residues involved in disulfide bonds, Zn-finger motifs or other post-translational modifications were eliminated from our datasets.

The data were then preprocessed using Chemical Computing Group (CCG) MOE software[46] using a custom SVL script. We have recently performed a comprehensive benchmark assessment of several methods for cysteine pK_a prediction.[28] The pK_a predictor tool in CCG MOE yielded the lowest mean absolute error among the methods tested for accurate cysteine pK_a prediction.[28] Protein structure preparation and refinement consisted of several tasks, including filling missing residues and loops within the structures, as well as restrained energy minimization following standard QuickPrep settings. After the protein structure preparation and refinement, we performed residue property calculations on the entire test set of structures. We computed four residue property descriptors in our calculations, namely: accessible surface area, percent exposure, residue charge, and pK_a. The resulting data was then parsed to extract properties relating to all cysteines (both modified and non-modified) in the dataset. We note that the covalent bonds between cysteine residues and ligands in the structures were broken prior to calculating residue descriptors, particularly pK_a.

Different protein preparation strategies were performed with the aim to improve the prediction accuracy for cysteine ligandability. We performed QuickPrep calculations at various stages in the structure preparation workflow for the entire dataset prior to computing residue property descriptors. **Scheme 1** provides a brief overview of the protein preparation workflow used in our study (see *Supporting Information* for details). We assessed the impact of different protein structural model choices on the predictive performance of the model descriptors, such as residue pK_a and percent solvent exposure. In particular, we assessed the differences between predictions on protein complexes left intact (chains together or 'CT') versus protein complexes split into component chains (chains apart or 'CA'). We also looked at whether running protein Quick Preparation — an

available tool from CCG MOE — at different times during the process impacted our model's predictive performance.

Scheme 1. Flowchart describing the protein preparation strategies used in our study for cysteine residue property prediction. *QuickPrep was run at various times in the protein preparation workflow prior to residue property prediction.*



We were also interested in whether a more targeted *in silico* approach to refining side chains (beyond QuickPrep) would result in higher-quality pK_a estimations, especially as we were observing cases in which sidechain orientations of residues neighboring cysteines in enzyme active sites were suboptimal. The orientation of the cysteine sidechain and its neighboring sidechains is crucial for accurate pK_a prediction. To address this concern, we performed sidechain conformer repacking and optimization with the Rosetta *fixbb* application in an attempt to correct poorly packed sidechain rotamers. Using a fixed backbone approximation, the neighborhood around each free and bound cysteine in each of the protein complexes was optimized using *fixbb*. We assessed neighborhood sizes of 7 Å, 8 Å, 9 Å, and 10 Å.

As a control and to analyze the impact of the different protein preparation strategies on the results, property predictions were also performed on the protein structures without running QuickPrep or *fixbb* (the **NQNC** dataset). Full datasets are available in the *Supporting Information*; the exact number of represented cysteines per dataset varies depending on the preparation method, as QuickPrep will build in missing cysteines if information is available.

Machine Learning Model Construction

Following the different structure preparation strategies, the data for the predicted residue descriptors were collected and split into test/train sets to build predictive machine learning (ML) models. We tested two types of ML models, namely: Logistic Regression (LR) and Random Forest (RF). We chose LR as the model of choice as it performed better than the RF model for our test case. Random seeds were set to ensure reproducibility of results. The model predicts the likelihood that a given class of cysteine in a target structure is covalently modifiable or not. An optimal cut-off value for the likelihood is calculated for each model based on the training set and is applied to the test set to attain a classification for each cysteine as either likely or unlikely to be covalently modifiable. We examined many different features of cysteine residues for the model-building exercise, including pK_a, solvent exposure, amino acid neighborhood, and electrostatic environment. Also, we adopted novel amino acid topological descriptors, T-Scale[47] and ST-Scale,[48] in our models to describe and capture the molecular structural information and residue local environment around the cysteines. Pocket analysis of cysteine locations in the protein structures was also incorporated as a feature in our ML model. The pocket finding software packages fpocket[49] and Site Finder[50] were used to predict potential protein-ligand binding sites in the structures.

In our ML model construction, we randomly selected 90% of the data for the training set and the remaining 10% for the test set. We utilized a random gene-wise split to assess whether prospective predictions would generalize to new genes and protein families. To ensure reproducibility, a set of random seeds was used to generate each test/train split. Because we were most interested in the correct prediction of true positives and the avoidance of false positives and false negatives in the context of using this model for drug discovery, F_1 score was used to evaluate the model performance. The reported F_1 score was the average across 10 random splits of the data for the different sets of covariates explored. We examined a variety of features to find the most parsimonious set of features that provided the best F_1 score on the test set. These features comprise an interpretable set of physicochemical aspects of a given cysteine residue that predict its liability to covalent reactivity.

RESULTS AND DISCUSSION

Cysteine pK_a, solvent exposure, local amino acid environment, and pocket inclusion correlate to the likelihood of cysteine modifiability

We began with a quantitative exploration of the CovPDB-derived dataset after removing protein systems without a properly covalently bound cysteine (e.g., those with cysteine bound to a metal), without any protein preparation or processing other than solvent, ion, and ligand removal. This set consisted of 886 complexes comprising 7246 cysteines, of which 1392 were covalently bound. We prioritized examining features that were likely to be physically and chemically meaningful. Prior models have often included cysteine pK_a and percent solvent exposure.[31] On average, ligandable cysteines tend to have a downshifted pK_a relative to unliganded (or free) cysteines (**Figure 2**). This observation has been used as the basis of many predictive models

for covalent modifiability of cysteine residues.[31] In the dataset that we examined, we also see a statistically significant difference between the pK_a of the covalently-bound cysteines and free cysteines. The observed pK_a range is broad, as would be expected based on the variability in experimental cysteine pK_a values.

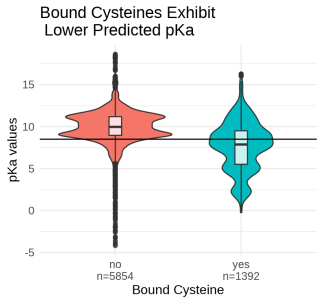


Figure 2. Computed pK_a for unbound (red) and bound (blue) cysteine residues. There is a downshift in predicted cysteine pK_a values for bound cysteines compared to the unbound ones. The horizontal black line indicates a pK_a of 8.6, which is the intrinsic or reference pK_a for cysteine residue in aqueous solution.[51]

The second major factor in most models is the percent solvent exposure: because cysteines that are not solvent exposed are unlikely to encounter (and thus bind to) a covalent inhibitor, it is reasonable to include solvent exposure as a model term that impacts likelihood of ligandability. In our datasets, covalently-modifiable cysteines are not linearly separable from free unliganded cysteines based on solvent exposure alone (**Figure 3**). Here, we use the log of the solvent exposure to increase sensitivity at the lower end of the percent solvent exposure scale, adding a small pseudo-value to avoid taking the log of 0. While covalently-modifiable cysteines are typically more solvent-exposed than their non-modifiable counterparts, there are a number of exposed cysteines that nonetheless are not covalently modified in the presence of a small molecule. This discrepancy may be due to the selectivity of the covalent binders in our datasets from COVPDB and CovBinderInPDB, as ligands available in these databases are unlikely to be indiscriminately binding small molecules due to the publication bias towards active, selective compounds. Training a model to predict relatively selective compounds, however, is an asset for the drug hunter as an indiscriminately binding small molecule is not likely to be very useful.

In addition to these two major factors, we were interested in exploring how the local amino acid environment and the inclusion of pocket information around cysteine residues affected our ability to accurately predict the ligandability of a cysteine.

The local amino acid environment is also important: it directly impacts the pK_a and may therefore serve as a reasonable replacement for pK_a given the difficulty of accurate *in silico* prediction of cysteine pK_a values.

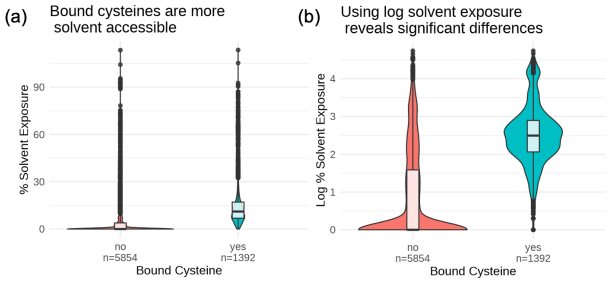


Figure 3. (a) Percent solvent exposure for unbound and bound cysteines. (b) Log percent solvent exposure for unbound and bound cysteines. Because of the log-normal distribution of solvent exposure percentages, the log transformation makes trends in the data clearer, exposing a clear difference in average exposure between bound and unbound. The transformed data also satisfy correlation test assumptions and yield a Pearson's *r* of 0.50.

We explored three different ways of encoding amino acid environment: the T-scale[47] of the local amino acids, the ST-scale[48] of the local amino acids, and frequency encoding of the local amino acids. For the purposes of our study, "local amino acids" were those within 8 Å of the sulfur of the cysteine residue. The T-scale and ST-scale encodings used the approach recommended in their papers. The frequency encoding was performed by counting the number of each unique amino acid in the neighborhood, giving a total of 20 new features for the model. If there were three alanines in the radius of the sulfur of the cysteine residue, for example, the value of the "A" feature would be 3.

While T-scales and ST-scales are intended to capture many attributes of the surrounding amino acids, and not just their number, we found better separation between our cases using the simple frequency encoding method rather than incorporating T-Scales or ST-Scales. The additional discriminative power of the model may be in part due to the greater number of associated features (20 features, in the case of frequency encoding; 5 or 8 features in the case of T-scales or ST-scales respectively), but the added interpretability of the model in using direct encoding of the amino acid environment is a benefit to the scientists using the model, and the number of parameters is still reasonable with respect to the number of training samples.

The final category of features considered was that of pocket inclusion. In these datasets, fpocket inclusion (**Table 1**) was a more discriminative criterion than SiteFinder inclusion (**Table 2**), even accounting for several different cut-offs for which detected pockets were included.

Table 1. FPocket pocket inclusion

	Cysteine bound?	
FPocket pocket detected?	yes	no
yes	**1279**	1023
no	113	**4831**

FPocket pocket detection gives a good true positive and true negative rate, with a Pearson's *r* of 0.63.

Table 2. SiteFinder Pocket inclusion

SiteFinder pocket detected?	Cysteine bound? yes	Cysteine bound? no
yes	**1320**	2129
no	72	**3725**

SiteFinder pocket detection gives a good true positive rate, but the false positive rate is much higher than FPocket. The correlation with cysteine binding status is concomitantly lower, with a Pearson's *r* of 0.46.

Protein preparation and sidechain repacking do not have a substantive impact on model performance

We explored different protein preparation strategies to investigate the impact of protein structural models on predicted residue properties and model performance. Using the QuickPrep application in CCG MOE, protein complexes were prepared to yield a good initial starting structure prior to residue property predictions. The QuickPrep calculations were performed at different stages in the structure preparation workflow for the entire dataset of protein complexes (**Scheme 1**). Both single protein chain and multimeric complexes were considered in the calculations. In addition to the QuickPrep calculations, sidechain repacking and optimization of residues around cysteines were performed via the Rosetta *fixbb* application. The *fixbb* was performed for residues 10 Å from the sulfurs of cysteines in the protein structures. Our results suggest that the different protein preparation strategies do not have a strong impact on the predictive performance of the model when considering different predictors such as pK_a alone, log solvent exposure alone, frequency AA encoding alone, and other models (**Figure 4**). We did observe that when chains are separated (the "chains apart" or CA datasets) at different times in the preparation, we do see that the log solvent exposure is a statistically significantly worse predictor of ligandability, indicating that chains-together (CT) is a better strategy. We hypothesize that many of the chains present in these structures may be found similarly bound *in vivo,* and so separating them may artificially inflate predicted cysteine solvent exposure.

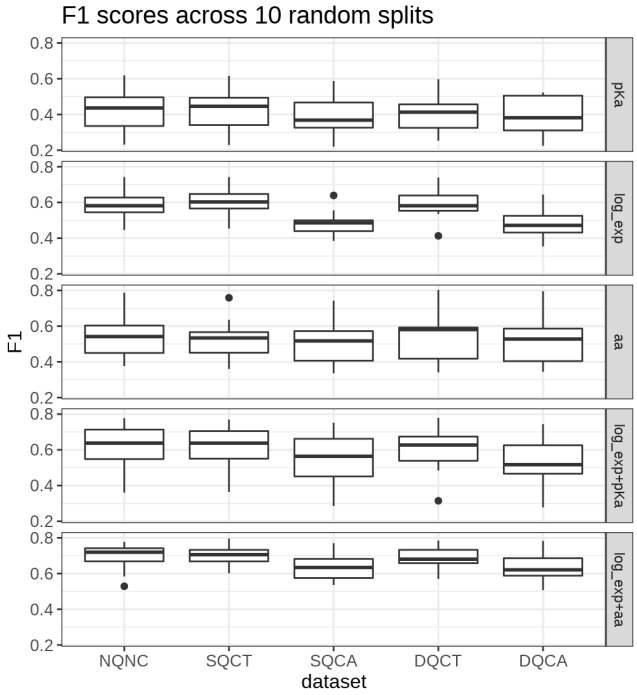


Figure 4. All model performance results are not significantly different than NQNC with the same covariates, with the exception of the SQCA/DQCA log_exp only model, which were statistically significantly worse ($p < 0.05$).

After visual inspection of several proteins, we wanted to assess whether repacking sidechains using Rosetta *fixbb* would result in better model performance. While QuickPrep does do some optimization of sidechains, it is not as comprehensive as the Rosetta repacking protocol. However, using Rosetta to repack sidechains within a 10 Å distance of each cysteine was not effective in raising the predictive performance over the NQNC baseline (**Figure 5**).

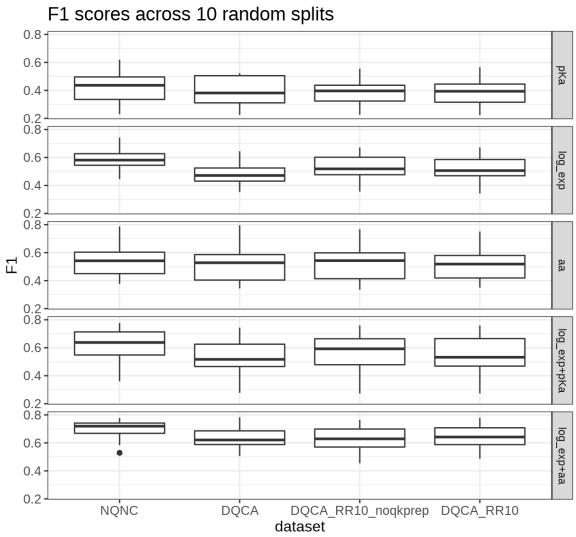


Figure 5. None of the models are statistically significantly different, except that DQCA has a significantly lower performance

than NQNC for the log_exp covariate model as shown in Figure 5.

Further, we were able to assess the differences in predicted pK_a and solvent exposure after Rosetta repacking, to elucidate some of the reason why prediction was not improved after theoretically improving the estimate of the local environment. We found that the differences in pK_a and solvent exposure clustered around zero, with noisy spread seemingly randomly distributed. These distributions were not different between bound and unbound cysteines (**Figure 6**).

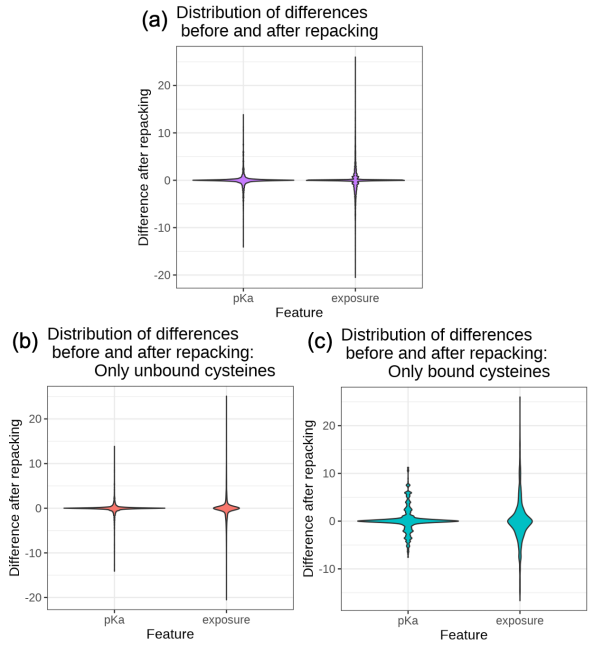


Figure 6. (a) Over all cysteines bound and unbound, the distribution of differences pre- and post-repacking is shown. (b) Distribution of differences for unbound cysteines only. (c) Distribution of differences for bound cysteines only.

With the lack of impact observed for both different protein preparation strategies as well as Rosetta sidechain repacking, we proceeded to use the NQNC dataset to build our final model. Our choice of model covariates was chosen based on a search of the potential space of model covariates (see Supporting Information); ultimately, it was found that a model that included the log solvent exposure, the presence of the cysteine in any fpocket-detected pocket, and the frequency encoded amino acid environment had the highest median performance across 10 random splits, giving an F1 score of 0.73. We call this model with the highest median performance **CovCysPredictor**.

A benefit of using an interpretable model such as logistic regression is that the coefficients can be examined to gain a better understanding of the model, and predictions can be traced back to specific contributions from different coefficients (**Table 3**). We observe that the coefficients attached to the twenty amino acids reveal some chemically-relevant trends: for example, the presence of a cysteine near the cysteine of interest decreases the probability of that cysteine being ligandable, possibly because of the presence of a disulfide bond between the two cysteines. Histidine is strongly positively correlated with the presence of a bound cysteine in the training set, which is sensible as it is often implicated in catalytic dyads/triads in cysteine-containing enzymatic processes. Glycine, the smallest amino acid, is positively correlated with ligandability, which may be due to the increased solvent exposure.

Table 3. Model coefficients of best-performing model

Intercept	-7.62		
log solv. exp.	1.18		
any fpocket	2.75	aa category	chemical justification
G	**0.55**	**special**	**solvent exposure**
A	0.31	hydrophobic	
V	-0.14	hydrophobic	
L	0.43	hydrophobic	
I	-0.22	hydrophobic	
P	-0.17	special	
M	-0.40	hydrophobic	
C	**-0.55**	**special**	**disulfide**
F	0.33	hydrophobic	
Y	0.51	hydrophobic	
W	1.12	hydrophobic	(see text)
H	**0.87**	**positive**	**catalytic dyad**
K	0.14	positive	
R	0.16	positive	
Q	-0.10	polar-uncharged	
N	0.24	polar-uncharged	
E	0.03	negative	
D	-0.61	negative	catalytic triad?
S	0.34	polar-uncharged	
T	0.01	polar-uncharged	

The coefficients of our logistic regression model reveal chemically relevant predictors of cysteine ligandability

Some of the coefficient directions, however, are less straightforward to interpret. Aspartic acid, for example, which is occasionally part of a catalytic triad with cysteine and histidine, is negatively correlated with ligandability. As part of the catalytic triad, it should be positively correlated with ligandability, otherwise the presence of the negative charge near cysteine is expected to increase its pKa leading to lower ligandability. Tryptophan is highly positively correlated with ligandability. These unexpected findings point to opportunities to understand the nature of ligandable cysteines better in future work.

Our simple, interpretable logistic regression model has comparable performance to other predictive ML models

Recent attempts to predict a cysteine's liability towards covalent modification have grown to include several tools based on more modern machine learning techniques, such as support vector machines,[31] graphical neural networks,[35] and decision tree-based models.[52] These tools represent gains in accuracy compared to earlier models but also have considerable increases in complexity.

Table 4. Comparison of our method to other structure-based cysteine ligandability prediction methods. *CovBinderInPDB's entries that were not present in CovPDB were used as a validation set

Method	Dataset	F1	AUPRC	Accuracy
CovCysPredictor (Ours)	CovPDB	0.73	**0.82**	**0.86**
	CovBinder-InPDB*	0.66	0.78	0.85
PriDeepCoSI	Du et al[35]		0.62	
DeepCoSI	Du et al[35]		0.76	
SVM	Zhang et al[31]			0.85

Unfortunately, there is no standard dataset or standard set of metrics used by methods for predicting cysteine ligandability. Uniformity and transparency of datasets used is critical for a fair assessment of available tools. However, in the case of Zhang et al,[31] it is unclear exactly what data was used as the website once used to host their data is unavailable. We therefore present their reported accuracy metric as a promising, but ultimately not definitive, comparison to our own model's performance on a different dataset. In the case of Du et al,[35] their external website used to host their training/testing data is also unavailable at the time of writing, although their training set is present in the supplemental material of their publication. Because their training and testing set has enormous overlap with our training set, it would not be fair to use their data as a test set for our algorithm.

Therefore, within the constraints of available data, we can show a side-by-side comparison of our models' area under the precision-recall curve (AUPRC) and accuracy with published results, which tentatively show our model to perform favorably (**Table 4**). Our model still performs favorably or equally well when tested on an orthogonal dataset (the PDBs in CovBinderInPDB that are not present in CovPDB). In addition, our model is considerably more interpretable than deep learning approaches and takes considerably less training time. We recognize the limitations of these analyses; any meaningful comparison between these tools is inherently limited by the lack of data availability and useful published benchmark sets with reliable, long-term storage solutions. We hope that the external datasets we have used, CovPDB and CovBinderInPDB, will become the benchmarks that tools use in the future. We make our data available in the Supporting Information attached to this manuscript as well as in GitHub.

Predictive power is partially maintained even when analysing *apo* versions of *holo* proteins from our dataset

The databases CovPDB and CovBinderInPDB both contain *holo* protein structures that contain the protein system as well as the covalently bound small molecule. In the ideal case, we would have trained on paired apo/holo datasets to create the most robust possible model. However, such a robust paired dataset does not yet exist. We were interested to see whether the predictive power of our model was nevertheless maintained even on *apo* protein structures, as scientists are more likely to find our tool useful if it has accurate predictions *before* any small molecule has been successfully covalently bound to the protein.

To assess the accuracy of CovCysPredictor on *apo* structures, we randomly selected a small set of proteins where a matched *holo/apo* pair were available. Then, the structures were cleaned of solvent, ions, cofactors, DNA, and RNA using open source tools (PyMOL,[53] Python/BioPython[54]) as we recommend in our GitHub for those who do not have access to MOE. The results in full are available in the Supporting Information.

We first assessed the performance of our model on the *holo* structures which were taken from our datasets (potentially seen in training). We found that 64% (7/11) of the *holo* structures in our holo/apo dataset were predicted correctly (**Table 5**). When assessing *apo* structures that our model had never seen before, 64% (7/11) of the *apo* structures were also predicted correctly. Using the strict criteria that both *holo* and *apo* versions of the protein are correctly predicted, we find a total success rate of around 36% (4/11, **Figure 7A**). However, assessing correctness here is nuanced. For example, for the pair 7b9m/6y58, the CYS45 that was known to be modifiable was not predicted to be modifiable; however, the rank order of the cysteines in the structures was correct, and the most likely cysteine was correctly identified as CYS45, even though the score fell below the threshold to be reported. Thus, the rank ordering was still reliable even though the absolute score was not.

Table 5. The 11 assessed holo/apo pairs, showing the known covalently-modifiable cysteine based on the holo structure, and the predicted ligandable cysteines for the holo and apo structures.

holo/apo pair	known cys	holo pred cys	apo pred cys
3zva/3zv8	147	147	147
5rfo/5r8t	145	145	145
7b9m/6y58	45	none	none
7brp/7bro	145	145;300	145;300
1uk4/1uk3	145	300	145;300
1a54/1a55	197	197	none
3v4o/3v55	464	none	464
1cte/1cpj	29	29;240	29;240
6yl1/6yl6	80	118	118
3o6t/3nof	37	37	none
1meg/1ppo	25	25	25

Another example of the ambiguity of "incorrect" assignations is found with 6yl1/6yl6 CDK2 structure where the gatekeeper phenylalanine is mutated to a cysteine (F80C) to enable covalent binding in the active site for inhibitor development. In this case, the predicted CYS118 is not the "ground-truth"

CYS80 based on 6yl1 structure, but CYS118 is also known to be covalently modifiable.[55]

A final example of the difficulty of ascertaining accuracy compared to ground truth is found with the 3o6t/3nof pair, where 3o6t (*holo*) is correctly predicted but 3nof (*apo*) does not predict any modifiable cysteines. However, the structure for 3nof contains a homodimer of the relevant protein whereas 3o6t contains a different heterocomplex. It would take knowledge of the biological system in question to be able to make the decision about whether the dimerization is likely to impact cysteine modification, but in this case, leaving the homodimer intact closed the cysteine off from being solvent accessible and therefore, it was predicted non-modifiable.

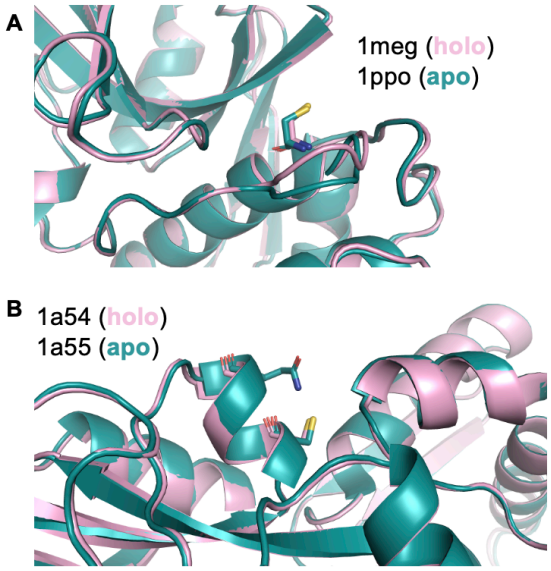


Figure 7. (**A**) A correctly-predicted *holo*/*apo* pair, 1meg/1ppo. (**B**) The positioning of the sidechain on GLN201 in the 1a55 structure (*apo* state) blocks access to the cysteine (C197) known to be ligandable in the *holo* state of this protein structure (1a54), and thus it is incorrectly predicted as unmodifiable in the *apo* form.

Broadly speaking, if we consider that a biologist running CovCysPredictor will be able to assess the appropriate state of a protein to be passed through the tool, all but one pair (10/11) was able to correctly (or at least reasonably) rank-order the cysteines for both *holo* and *apo* structures. The pair 1a54/1a55 was correctly predicted for the *holo* partner, but not the *apo* one, due to what appears to be an induced pocket which does not exist in the *apo* structure, but does exist in the *holo* structure when the sidechain of GLN201 does not occlude CYS197 (**Figure 7B**). Cryptic/induced pockets and structural flexibility in general are difficult for CovCysPredictor to handle, as it learns and predicts on given static structures. In the future, if a paired *holo*/*apo* dataset were available for training, factors such as the cysteine residue environment and the sulfur distance to the protein surface in both *apo* and corresponding *holo* states could be included as a feature to improve the likelihood of being able to predict the possibility of ligand-induced binding events of targetable cysteine residues.

CONCLUSIONS

In summary, we have examined two major recent protein structural databases (CovPDB and CovBinderInPDB) for developing a machine learning model to predict ligandable cysteine sites in the proteome. Our approach involved training and testing an interpretable machine learning model to identify cysteines which are liable to be covalently modified in a selective fashion. We explored the inclusion or exclusion of different physicochemical features in our models, including residue pK_a, solvent exposure, and residue electrostatics environment. In addition, we employed descriptors from protein-ligand pocket detection algorithms, fpocket and SiteFinder, as features in our machine learning model to differentiate between purely reactive cysteines, which may be highly surface exposed, and "ligandable" cysteines, which would ideally be placed in a pocket that could be selectively targeted by a small molecule drug. In developing our final logistic regression model, we also explored several protein preparation strategies and their impact on model predictive performance. Our resulting logistic regression model achieved comparable performance to other predictive more complex ML models, yielding an F_1 performance of 0.73. Moving forward, we envision this model to provide valuable insights to guide early cysteine-targeting covalent drug design strategies. Overall, our work demonstrates the potential of interpretable machine learning algorithms to impact efforts in early drug discovery.

DATA AND SOFTWARE AVAILABILITY

Additional data and information about model performance for the ML techniques explored are available in the Supporting Information. All code and generated data files are publicly available at Bryn-MarieR/CovCysPredictor (github.com).

ASSOCIATED CONTENT

Supporting Information

The Supporting Information includes details of the datasets used for the machine learning models and further explanation of the preparation protocols.

The Supporting Information is available free of charge on the ACS Publications website.

AUTHOR INFORMATION

Authors

Bryn Marie Reimer* – Novartis Biomedical Research, 181 Massachusetts Avenue, Cambridge, MA 02139, United States.
Ernest Awoonor-Williams – Novartis Biomedical Research, 181 Massachusetts Avenue, Cambridge, MA 02139, United States;
Present address: Schrödinger Inc., New York, NY 10036, United States

Andrei A. Golosov – Novartis Biomedical Research, 181 Massachusetts Avenue, Cambridge, MA 02139, United States
Viktor Hornak* – Novartis Biomedical Research, 181 Massachusetts Avenue, Cambridge, MA 02139, United States

* Co-corresponding

Author Contributions

The manuscript was written through the contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding Sources

A.A.G. is a current employee of Novartis Biomedical Research, who funded this work. B.M.R., E.A.-W., and V.H. are former employees of Novartis Biomedical Research.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors would like to acknowledge William Long, from Chemical Computing Group (CCG), for SVL script contributions and useful discussion. They would also like to thank David Thompson from CCG for early discussion about MOE databases and SVL customization. E.A-W acknowledges Lynn McGregor, Dave Barkan, and Anne Granger, as well as support from the Innovation Postdoctoral Fellowship Program at Novartis Biomedical Research. We thank the reviewers for providing valuable constructive feedback, which helped to improve the quality of the manuscript.

ABBREVIATIONS

BTK, Bruton's tyrosine kinase; CCG, chemical computing group; Cys, cysteine; LR, logistic regression; RF, random forest; ML, machine learning; PCA, principal component analysis; MOE, molecular operating environment; PDB, protein data bank; TCI, targeted covalent inhibitor.

REFERENCES

(1) Singh, J.; Petter, R. C.; Baillie, T. A.; Whitty, A. The Resurgence of Covalent Drugs. *Nat. Rev. Discov.* **2011**, *10*, 307–317.
(2) Baillie, T. A. Targeted Covalent Inhibitors for Drug Design. *Angew. Chemie - Int. Ed.* **2016**, *55*, 13408–13421.
(3) Awoonor-Williams, E.; Walsh, A. G.; Rowley, C. N. Modeling Covalent-Modifier Drugs. *Biochim. Biophys. Acta - Proteins Proteomics* **2017**, *1865*, 1664–1675.
(4) Abdeldayem, A.; Raouf, Y. S.; Constantinescu, S. N.; Moriggl, R.; Gunning, P. T. Advances in Covalent Kinase Inhibitors. *Chem. Soc. Rev.* **2020**, *49*, 2617–2687.
(5) Vita, E. De. 10 Years into the Resurgence of Covalent Drugs. *Future Med. Chem.* **2021**, *13*, 193–210.
(6) Boike, L.; Henning, N. J.; Nomura, D. K. Advances in Covalent Drug Discovery. *Nat. Rev. Drug Discov.* **2022**, *21*, 881–898.
(7) Singh, J. The Ascension of Targeted Covalent Inhibitors. *J. Med. Chem.* **2022**, *65*, 5886–5901.
(8) Awoonor-Williams, E.; Abu-Saleh, A. A.-A. A. Covalent and Non-Covalent Binding Free Energy Calculations for Peptidomimetic Inhibitors of SARS-CoV-2 Main Protease. *Phys. Chem. Chem. Phys.* **2021**, *23*, 6746–6757.
(9) Sutanto, F.; Konstantinidou, M.; Dömling, A. Covalent Inhibitors: A Rational Approach to Drug Discovery. *RSC Med. Chem.* **2020**, *11*, 876–884.
(10) Pace, N. J.; Weerapana, E. Diverse Functional Roles of Reactive Cysteines. *ACS Chem. Biol.* **2013**, *8*, 283–296.
(11) Giles, N. M.; Giles, G. I.; Jacob, C. Multiple Roles of Cysteine in Biocatalysis. *Biochem. Biophys. Res. Commun.* **2003**, *300*, 1–4.
(12) Leproult, E.; Barluenga, S.; Moras, D.; Wurtz, J. M.; Winssinger, N. Cysteine Mapping in Conformationally Distinct Kinase Nucleotide Binding Sites: Application to the Design of Selective Covalent Inhibitors. *J. Med. Chem.* **2011**, *54*, 1347–1355.
(13) Liu, Q.; Sabnis, Y.; Zhao, Z.; Zhang, T.; Buhrlage, S. J.; Jones, L. H.; Gray, N. S. Developing Irreversible Inhibitors of the Protein Kinase Cysteinome. *Chem. Biol.* **2013**, *20*, 146–159.
(14) Visscher, M.; Arkin, M. R.; Dansen, T. B. Covalent Targeting of Acquired Cysteines in Cancer. *Curr. Opin. Chem. Biol.* **2016**, *30*, 61–67.
(15) Maurais, A. J.; Weerapana, E. Reactive-Cysteine Profiling for Drug Discovery. *Curr. Opin. Chem. Biol.* **2019**, *50*, 29–36.
(16) Awoonor-Williams, E.; Rowley, C. N. Modeling the Binding and Conformational Energetics of a Targeted Covalent Inhibitor to Bruton's Tyrosine Kinase. *J. Chem. Inf. Model.* **2021**, *61*, 5234–5242.
(17) Lu, X.; Smaill, J. B.; Patterson, A. V.; Ding, K. Discovery of Cysteine-Targeting Covalent Protein Kinase Inhibitors. *J. Med. Chem.* **2022**, *65*, 58–83.
(18) Awoonor-Williams, E. Modelling Covalent Modification of Cysteine Residues in Proteins, Memorial University of Newfoundland, 2020.
(19) Awoonor-Williams, E. Estimating the Binding Energetics of Reversible Covalent Inhibitors of the SARS-CoV-2 Main Protease: An in Silico Study. *Phys. Chem. Chem. Phys.* **2022**, *24*, 23391–23401.
(20) Tao, Y.; Remillard, D.; Vinogradova, E. V.; Yokoyama, M.; Banchenko, S.; Schwefel, D.; Melillo, B.; Schreiber, S. L.; Zhang, X.; Cravatt, B. F. Targeted Protein Degradation by Electrophilic PROTACs That Stereoselectively and Site-Specifically Engage DCAF1. *J. Am. Chem. Soc.* **2022**, *144*, 18688–18699.
(21) Awoonor-Williams, E.; Rowley, C. N. How Reactive Are Druggable Cysteines in Protein Kinases? *J. Chem. Inf. Model.* **2018**, *58*, 1935–1946.
(22) Roos, G.; Foloppe, N.; Messens, J. Understanding the PK(a) of Redox Cysteines: The Key Role of Hydrogen Bonding. *Antioxid. Redox Signal.* **2013**, *18*, 94–127.
(23) Marino, S. M.; Gladyshev, V. N. Analysis and Functional Prediction of Reactive Cysteine Residues. *J. Biol. Chem.* **2012**, *287*, 4419–4425.
(24) Awoonor-Williams, E.; Rowley, C. N. Evaluation of Methods for the Calculation of the PKa of Cysteine Residues in Proteins. *J. Chem. Theory Comput.* **2016**, *12*, 4662–4673.
(25) Pahari, S.; Sun, L.; Alexov, E. PKAD: A Database of Experimentally Measured PKa Values of Ionizable Groups in Proteins. *Database* **2019**, *2019*.
(26) Bulaj, G.; Kortemme, T.; Goldenberg, D. P. Ionization−Reactivity Relationships for Cysteine Thiols in Polypeptides †. *Biochemistry* **1998**, *37*, 8965–8972.
(27) 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.
(28) Awoonor-Williams, E.; Golosov, A. A.; Hornak, V. Benchmarking In Silico Tools for Cysteine p K a Prediction. *J. Chem. Inf. Model.* **2023**.
(29) Soylu, İ.; Marino, S. M. Cy-Preds : An Algorithm and a Web Service for the Analysis and Prediction of Cysteine Reactivity. *Proteins Struct. Funct. Bioinforma.* **2016**, *84*, 278–291.
(30) Zhang, Y.; Zhang, D.; Tian, H.; Jiao, Y.; Shi, Z.; Ran, T.; Liu, H.; Lu, S.; Xu, A.; Qiao, X.; Pan, J.; Yin, L.; Zhou, W.; Lu, T.;

Chen, Y. Identification of Covalent Binding Sites Targeting Cysteines Based on Computational Approaches. *Mol. Pharm.* **2016**, *13*, 3106–3118.

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

(32) Zhao, Z.; Liu, Q.; Bliven, S.; Xie, L.; Bourne, P. E. Determining Cysteines Available for Covalent Inhibition Across the Human Kinome. *J. Med. Chem.* **2017**, *60*, 2879–2889.

(33) 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.

(34) 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.

(35) 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*.

(36) White, M. E. H.; Gil, J.; Tate, E. W. Proteome-Wide Structural Analysis Identifies Warhead- and Coverage-Specific Biases in Cysteine-Focused Chemoproteomics. *Cell Chemical Biology* **2023**, *30*, pp P828–838.

(37) Awoonor-Williams, E.; Isley, W. C.; Dale, S. G.; Johnson, E. R.; Yu, H.; Becke, A. D.; Roux, B.; Rowley, C. N. Quantum Chemical Methods for Modeling Covalent Modification of Biological Thiols. *J. Comput. Chem.* **2020**, *41*, 427–438.

(38) Awoonor-Williams, E.; Kennedy, J.; Rowley, C. N. Measuring and Predicting Warhead and Residue Reactivity. In *Annual Reports in Medicinal Chemistry*; Ward, R. A., Grimster, N. P., Eds.; Elsevier Inc., 2021; pp 203–227.

(39) Wang, H.; Chen, X.; Li, C.; Liu, Y.; Yang, F.; Wang, C. Sequence-Based Prediction of Cysteine Reactivity Using Machine Learning. *Biochemistry* **2018**, *57*, 451–460.

(40) Marino, S. M.; Salinas, G.; Gladyshev, V. N. Computational Functional Analysis of Cysteine Residues in Proteins. In *Redox Chemistry and Biology of Thiols*; Elsevier, 2022; pp 59–80.

(41) Weerapana, E.; Wang, C.; Simon, G. M.; Richter, F.; Khare, S.; Dillon, M. B. D.; Bachovchin, D. A.; Mowen, K.; Baker, D.; Cravatt, B. F. Quantitative Reactivity Profiling Predicts Functional Cysteines in Proteomes. *Nature* **2010**, *468*, 790–797.

(42) 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.

(43) 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.* **2021**, No. 23, 1–6.

(44) Guo, X.-K.; Zhang, Y. CovBinderInPDB: A Structure-Based Covalent Binder Database. *J. Chem. Inf. Model.* **2022**, *62*, 6057–6068.

(45) Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(46) Molecular Operating Environment (MOE), 2022.02 Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910. Montreal, QC H3A 2R7, Canada, 2022.

(47) Tian, F.; Zhou, P.; Li, Z. T-Scale as a Novel Vector of Topological Descriptors for Amino Acids and Its Application in QSARs of Peptides. *J. Mol. Struct.* **2007**, *830*, 106–115.

(48) Yang, L.; Shu, M.; Ma, K.; Mei, H.; Jiang, Y.; Li, Z. ST-Scale as a Novel Amino Acid Descriptor and Its Application in QSAM of Peptides and Analogues. *Amino Acids* **2010**, *38*, 805–816.

(49) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinformatics* **2009**, *10*, 168.

(50) Soga, S.; Shirai, H.; Kobori, M.; Hirayama, N. Use of Amino Acid Composition to Predict Ligand-Binding Sites. *J. Chem. Inf. Model.* **2007**, *47*, 400–406.

(51) Thurlkill, R. L.; Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. PK Values of the Ionizable Groups of Proteins. *Protein Sci.* **2006**, *15*, 1214–1218.

(52) Liu, R.; Clayton, J.; Shen, M.; Bhatnagar, S.; Shen, J. K. Machine Learning Models to Interrogate Proteome-Wide Covalent Ligandabilities Directed at Cysteines. *JACS Au.* **2024**, 4, 4, 1374–1384

(53) The PyMOL Molecular Graphics System. 3.0 Schrödinger, LLC.

(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) Craven, G. B., Affron, D. P., Allen, C. E., Matthies, S., Greener, J. G., Morgan, R. M. L., Tate, E. W., Armstrong, A., Mann, D. J. High-Throughput Kinetic Analysis for Target-Directed Covalent Ligand Discovery. *Angewandte Chemie (International ed. in English)* **2009,** *57*(19), 5257–5261.

Table of Contents Graphic