TopCysteineDB: A Cysteinome-wide Database Integrating Structural and Chemoproteomics Data for Cysteine Ligandability Prediction ☆

Michele Bonus [1,†], Julian Greb [1,†], Jaimeen D. Majmudar [2], Markus Boehm [2], Magdalena Korczynska [2], Azadeh Nazemi [2], Alan M. Mathiowetz [2], and Holger Gohlke [1,3,*]

1 - ***Institute for Pharmaceutical and Medicinal Chemistry,*** *Heinrich Heine University, Düsseldorf 40225 Düsseldorf, Germany*
2 - ***Pfizer Research & Development,*** *Cambridge, MA 02139, United States*
3 - ***Institute of Bio- and Geosciences (IBG4: Bioinformatics),*** *Forschungszentrum Jülich GmbH, 52425 Jülich, Germany*

Correspondence to Holger Gohlke: Institute for Pharmaceutical and Medicinal Chemistry, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany. *gohlke@hhu.de, h.gohlke@fz-juelich.de (H. Gohlke)*
https://doi.org/10.1016/j.jmb.2025.169196
Edited by Michael Sternberg

Abstract

Development of targeted covalent inhibitors and covalent ligand-first approaches have emerged as a powerful strategy in drug design, with cysteines being attractive targets due to their nucleophilicity and relative scarcity. While structural biology and chemoproteomics approaches have generated extensive data on cysteine ligandability, these complementary data types remain largely disconnected. Here, we present *TopCysteineDB*, a comprehensive resource integrating structural information from the PDB with chemoproteomics data from activity-based protein profiling experiments. Analysis of the complete PDB yielded 264,234 unique cysteines, while the proteomics dataset encompasses 41,898 detectable cysteines across the human proteome. Using *TopCovPDB*, an automated classification pipeline complemented by manual curation, we identified 787 covalent cysteines and systematically categorized other functional roles, including metal-binding, cofactor-binding, and disulfide bonds. Mapping residue-wise structural information to sequence space enabled cross-referencing between structural and proteomics data, creating a unified view of cysteine ligandability. For *TopCySPAL*, a machine learning model was developed, integrating structural features and proteomics data, achieving strong predictive performance (AUROC: 0.964, AUPRC: 0.914) and robust generalization to novel cases. *TopCysteineDB* and *TopCySPAL* are freely accessible through a webinterface, *TopCysteineDBApp* (https://topcysteinedb.hhu.de/), designed to facilitate exploration of cysteine sites across the human proteome. The interface provides an interactive visualization featuring a color-coded mapping of chemoproteomics data onto cysteine site structures and the highlighting of identified peptide sequences. It offers customizable dataset downloads and ligandability predictions for user-provided structures. This resource advances targeted covalent inhibitor design by providing integrated access to previously dispersed data types and enabling systematic analysis and prediction of cysteine ligandability.

☆ This article is part of a special issue entitled: 'Computation Resources (2025)' published in Journal of Molecular Biology.

Introduction

In recent years, the development of targeted covalent inhibitors (TCIs) has experienced a renaissance in drug design [1], offering unique advantages, including enhanced potency, prolonged target engagement, and the potential for improved selectivity [2–5]. These small-molecule inhibitors employ electrophilic warheads – functional groups such as acrylamides – that irreversibly or reversibly form covalent bonds with nucleophilic amino acid residues in proteins, thereby modulating their function [6,7].

Among the nucleophilic residues, cysteine residues stand out due to their unique reactivity and relatively low abundance [8]. Particularly in oncology, cysteine-focused covalent inhibitors have achieved remarkable clinical success by targeting proteins previously considered undruggable [9], including sotorasib, which covalently binds the $\text{KRAS}^{\text{G12C}}$ mutant [10]. The design of such covalent drugs critically depends on the strategic selection of suitable target cysteine residues, which in turn requires a profound understanding of their ligandability. Consequently, experimental and computational methods, as well as data resources that facilitate the assessment of cysteine ligandability, are of significant interest in medicinal chemistry.

The advent of chemical proteomics, particularly the development of activity-based protein profiling (ABPP), has revolutionized the global assessment of cysteine ligandability in a cellular context [11]. In ABPP, cell or lysate samples are treated with reactive scout molecules or libraries of warhead-carrying covalent fragments. By subsequent reaction with a pan-cysteine-reactive iodoacetamide probe and comparison with an untreated sample, reactive cysteines can be identified across complex proteomes [12–15]. This methodology has also been adapted for proteome-wide covalent ligand discovery using fragment libraries [16]. The results of such studies have recently been compiled into the CysDB database, offering an overview of the quantitative chemoproteomics data of the human cysteinome [17].

Interpreting the functional relevance of cysteine engagement identified through chemoproteomics requires distinguishing several related concepts. The interaction of a cysteine with a pan-reactive probe like iodoacetamide primarily reflects its accessible *reactivity* – its intrinsic chemical propensity, modulated by the local microenvironment (e.g., pK_a). However, when screening large libraries of diverse electrophiles [12–14], observing selective engagement of a cysteine by only a subset of specific fragments provides strong experimental evidence for *ligandability*. Ligandability goes beyond mere reactivity, incorporating the presence and accessibility of a suitable binding pocket capable of accommodating a ligand; a reactive cysteine may not be ligandable if it lacks such a pocket. Ultimately, the goal in drug development is to identify *druggable* sites. Druggability builds upon ligandability, further requiring that modulation of the target protein via ligand binding yields a therapeutic benefit with an acceptable safety profile, often implying the pocket has characteristics amenable to binding optimized, drug-like molecules. Predicting ligandability, by integrating evidence for reactivity with structural context, is therefore a key step towards identifying druggable covalent targets. Several specialized databases have been developed for covalent drug discovery. The *CovPDB* [18] catalogs structural information from covalent protein–ligand complexes, while *CovalentInDB* [19] focuses on covalent inhibitors and their properties, including protein–ligand interactions in its most recent version [20]. Other databases include the *Cysteinome* [21], *cBinderDB* [22], and *CovBinderInDB* [23].

Additionally, computational prediction methods for cysteine ligandability are invaluable to covalent drug design. Classic physics-based methods like molecular dynamics simulations have been used to determine cysteine pK_a values [24,25]. However, these methods require expert users and are computationally demanding. Consequently, a variety of data-driven approaches, including modern supervised machine learning (ML) algorithms, have been applied. These differ in the form of the required input, the features utilized for prediction, the algorithm employed, and the source of the cysteine ligandability labels. Soylu et al. combined a sequence-based approach with energetic hydrogen-bond interaction analysis of cysteine sites to develop Cy-preds [26]. Wang et al. developed *sbPCR*, a purely sequence-based support vector machine (SVM) model to predict hyperreactive cysteines based on isotopic tandem orthogonal proteolysis (isoTOP)-ABPP-derived ligandability data [27]. In contrast, Zhang et al. developed an SVM model based on structural data, incorporating features such as solvent accessibility and predicted pK_a values combined with structure-based labels of cysteine reactivity [28]. Recently, increasingly sophisticated ML architectures emerged. *HyperCys* introduced a stacked model approach combining structural parameters generated from *CovPDB* data with sequence data [29], while *DeepCoSI* implemented a graph convolutional network architecture trained on *CovalentInDB* structural data and pocket information [30]. The Shen group introduced tree-based and convolutional neural network (CNN) models that leveraged structural data and descriptors [31]. Recently, *DrugMap* has integrated large-scale isoTOP-ABPP data with CNN-based prediction models that use structural descriptors as well [15]. Other models aim to predict reactivity changes influenced by specific factors, such as post-translational modifications [32]. Further structure-based machine learning approaches include interpretable models trained on covalent databases (*CovCysPredictor* [33]) and random forests using integrated proteomics and structural data to predict reactivity towards specific probes like IAA (*CIAA* [34]).

While recent efforts integrate structural and proteomics data for specific reactivity predictions [34], a resource offering systematically annotated, PDB-wide structural classifications combined with multiple large-scale chemoproteomics datasets to facilitate broad structure- and ML-based ligandability prediction on the human cysteinome-scale is still needed. Here, we present *TopCysteineDB*, a resource linking these two essential data types, and use it to establish a cysteine structure-ligandability relationship that improves our ability to predict covalent ligandability. Initially, we performed a nuanced, PDB-wide classification of covalent cysteine site structures using *TopCovPDB*, an automated algorithm analyzing the local residue environment, complemented by extensive manual curation. This resulted in a high-quality dataset of covalently modified cysteines. This approach is more comprehensive than specialized databases like *CovPDB*, yet it effectively distinguishes highly relevant structural evidence for ligandability, e.g., TCI–protein complexes, from less relevant structures, such as artifact- or cofactor-bound cysteines, which may not be resolved in less detailed datasets. Next, mapping the residue-wise structural information to the UniProt sequence space enabled the interconnection of all structural data for each unique cysteine. This yielded a global classification of a cysteine's ligandability, integrated across all experimentally observed structural states. This way, the structural flexibility of the residue environment and especially changes upon covalent engagement were captured. Additionally, the mapping allows for cross-referencing ligandability data from extensive chemoproteomics experiments with structural information, providing further relevant and comparable ligandability data for the human cysteinome. By integrating these complementary data types, our approach enhances the identification of patterns associated with cysteine ligandability. Consequently, predictive models generated using the resulting database are expected to align well with factors determining covalent druggability.

Accordingly, we developed the ML model *TopCySPAL* (Cysteine Structure–Proteomics-Augmented Ligandability predictor) that followed a combined label strategy based on structural and ABPP ligandability data, with a particular focus on label quality and accounting for uncertainty in negative labels. Here, the structure-based features used were a combination of structural descriptors of the cysteine environment and state-of-the-art SaProt residue embeddings [35]; SaProt residue embeddings are vector representations of individual amino acid residues in proteins enabling the capture of residue-level biochemical and spatial features for downstream tasks. *TopCysteineDB* is accessible via an interactive web interface featuring a uniquely integrative visualization of the available ligandability data mapped onto classified protein structures as well as the predictive capabilities of *TopCySPAL*.

Results & Discussion

TopCysteineDB: A data resource that links structural and proteomics data

Integration of structural and proteomics data for a comprehensive cysteine ligandability resource. *TopCysteineDB* integrates two complementary data types: experimental structural information from the PDB and chemoproteomics data from three large-scale isoTOP-ABPP studies [12–14] (Figure 1A, Figure 1B). To enable a systematic analysis of cysteine ligandability across these experiments, an SQL database was designed that efficiently cross-references each datapoint to a unique cysteine in the UniProt sequence space. For the structure dataset, the PDB and UniProt residue numberings were mapped using residue-level SIFTS data [36,37]. Analyzing the entire PDB yielded 192,716 structures, corresponding to 264,234 unique cysteines (Figure 1C). Of these, 58,505 structures were of human origin covering 58,494 cysteines across 7,631 unique human proteins, indicating that currently 22% of the human cysteinome (58,494 out of a total of 262,566 cysteines identified in the reference proteome) is structurally resolved. The proteomics dataset encompasses 41,898 detectable human cysteines (16% cysteinome coverage) across 9,319 unique proteins (45% of the human proteome) based on ~4.5 million competition ratio (CR) measurements from three comprehensive isoTOP-ABPP studies [12–14]. Of these, 9,427 cysteines were identified as "ABPP-ligandable" (Competition Ratio (CR) $\geq$ 4). While current coverage differs from larger resources like *CysDB* (62,888 cysteines) or *DrugMap* (78,523 cysteines), this database substantially covers the druggable human cysteinome combined with closely integrated structural information, featuring a reliable mapping between experimental cysteine site structures and UniProt sequence space. *TopCysteineDB*'s modular architecture enables continuous integration of new proteomics datasets, ensuring high-quality data curation.

Structure-based classification pipeline enables systematic analysis of cysteine-partner interactions. To establish a comprehensive understanding of cysteine ligandability across the available experimental structural data, we developed *TopCovPDB*, an automated classification pipeline that systematically categorizes cysteines based on their interaction partners in protein structures. The pipeline analyzes structural contacts to identify distinct types of cysteine site structures/modifications,

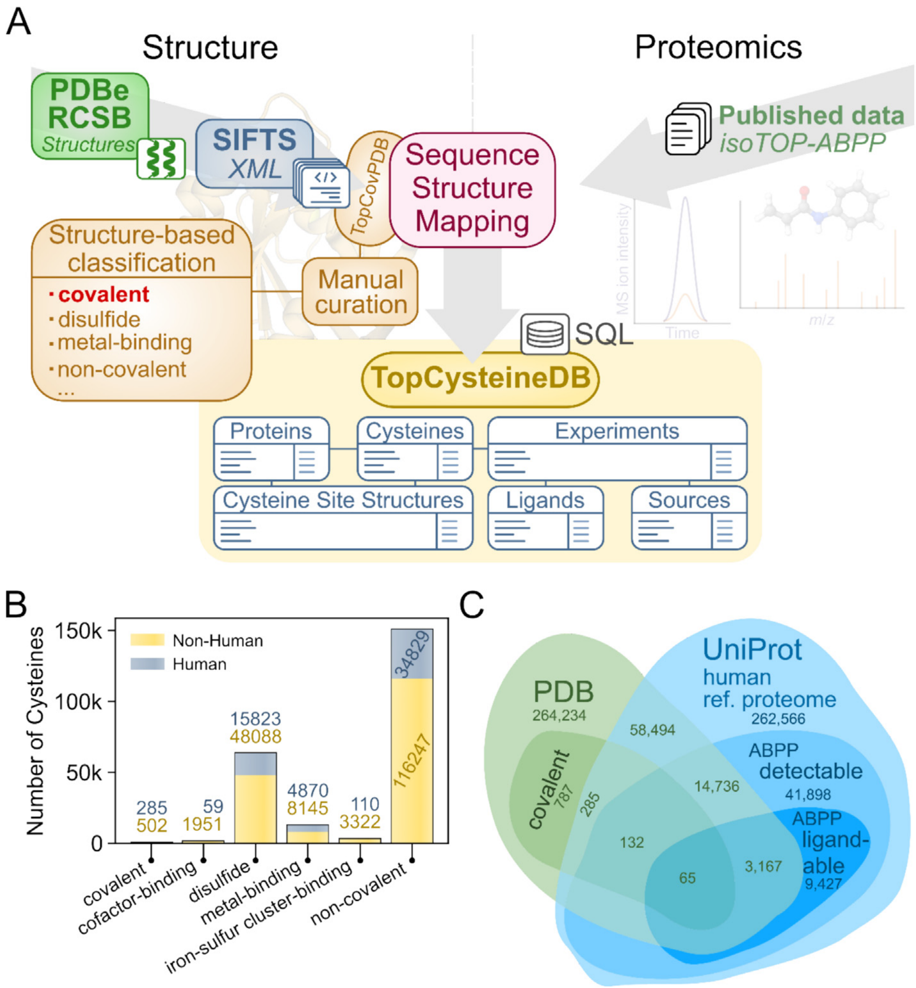


Figure 1. ***TopCysteineDB***. (**A**) Workflow for the generation of *TopCysteineDB*, a relational SQL database integrating cysteine ligandability evidence from the PDB and *iso*-TOP-ABPP experiments by mapping comprehensive annotated structural data (*TopCovPDB*) to the UniProt sequence space. See text for details. (**B**) Number of unique cysteine sites by structural classifications based on the *TopCovPDB workflow* after manual curation. (**C**) Venn diagram depicting the overlap of unique cysteines included in the database: with a PDB entries (light green), with covalent ligand-complex structures (green), belonging to the human cysteinome (light blue, part of the human reference proteome), detected by ABPP experiments (blue), and identified as ABPP-ligandable (dark blue, CR ≥ 4).

including disulfide bonds (63,911), metal-binding (13,015), iron-sulfur cluster-binding (3,432), and cofactor-binding (2,010) cysteines (Figure 1B). Importantly, cysteines that form covalent bonds with partners that do not belong to these predefined categories are classified as covalent, typically indicating interactions with covalent small-molecules. In the absence of any covalent interaction partner, cysteines are classified as non-covalent. To ensure the high quality of these classifications, particularly for the covalent class of cysteines crucial for drug design applications, 9,175 automatically generated classifications were reviewed in a thorough manual curation. This curation was essential to verify that cysteines classified as covalent were not binding to cofactors, metal ions, or iron-sulfur clusters and to resolve ambiguous cases with less definitive interaction geometry.

For comprehensive coverage of the structural cysteine space, we opted not to apply the rigorous thresholds utilized in *CovPDB* (e.g., resolution ≤ 2.5 Å) [18]. The *TopCovPDB* classification revealed 787 covalent cysteines, 285 of human origin across 692 unique proteins (235 human proteins). While this number exceeds the 309 unique cysteines in *CovPDB* (115 human cysteines; across 290 proteins, 91 of human origin), it remains lower than the 1,133 cysteines (778 proteins) reported in *LigCys3D* [31], likely due to our more stringent manual curation and fine-grained distinction between covalent sites and other modifications (Figure 1B). Importantly, our dataset additionally includes 151,076 non-covalent cysteines

(36,959 proteins), providing negative dataset examples crucial for ML applications.

Cross-validation. The overlap between structural and proteomics data provides valuable opportunities for cross-validation. A total of 14,736 cysteines (3,841 proteins) are structurally resolved and detectable in proteomics experiments. For 3,167 cysteines (1,648 proteins) there is experimental evidence of ligandability. 132 cysteines (117 proteins) are structurally classified as covalent and detectable in proteomics experiments, with 65 of these (63 proteins) showing experimental evidence of ligandability (Figure 1C).

The machine learning model TopCySPAL enables accurate ligandability prediction

To predict potentially ligandable cysteines, we developed the ML model *TopCySPAL* that integrates structural features with proteomics data. While recent approaches have shown promise, they typically rely on single data modalities. For instance, the *LigCys3D* database and the corresponding ML model *DeepCys* [31] employ purely structure-based labels. In contrast, our approach creates a more integrated training dataset by interconnecting structural and proteomics data in *TopCysteineDB*: Positive labels were assigned to unique cysteines either classified as covalent based on structural evidence (classified covalent in any structure) or by showing chemoproteomics evidence of ligandability; negative labels were only assigned when other unique cysteines in the same protein demonstrated either structural evidence of covalent binding or were detectable in proteomics experiments, indicating that all cysteines have been analyzed in a ligandability experiment (structural or proteomics-based) but the cysteine of interest was not found to be ligandable. Finally, for detectable cysteines lacking experimental structural information, the dataset has been supplemented with AlphaFoldDB [38] structures.

The initial dataset comprised 343,722 samples (47,221 positive, 296,501 negative), derived from proteins with both structural characterization and proteomics data. Protein structures were grouped by AlphaFoldDB [38] clusters [39] to prevent homology-based data leakage, ensuring that all structures of a unique cysteine were kept in a single split/fold (see SI section 1.2 for details). To prevent bias from overrepresented proteins, we limited each unique cysteine to 32 structural instances, resulting in 17,062 positive and 152,203 negative samples. While Shen et al. [31] explicitly analyzed conformational variability by binning structures according to solvent-accessible surface area (SASA) values (9,992 positive structures from 1,133 unique cysteines and 10,267 negative structures from 3,084 unique cysteines), our sampling approach with multiple structural instances similarly captures structural diversity while avoiding overrepresentation. To address the inherent uncertainty in negative labels, we employed a Positive-Unlabeled (PU) learning framework [40] for learning classifiers from fewer labeled positive examples with many uncertain negative ones, which resulted in a final training set of 17,062 positive examples from 1,465 unique cysteines and 76,102 high-confidence negative examples from 8,444 unique cysteines.

For each cysteine site structure, SaProt embeddings [35] and a set of structural descriptors (such as SASA, secondary structure classifications, and neighboring residue atom types, counts, and distances), calculated using Biotite [41,42] and DSSP [43] (see SI Section 1.2.1.1 for details), were generated, on which jointly an XGBoost model was trained (Figure 2A). Subsequent recursive feature elimination using SHAP values [44] identified an optimal set of 196 features (160 embedding dimensions, 36 structural descriptors) that maintained 92% of the full model performance while further preventing overfitting by an 86% reduction in feature dimensionality. The final model *TopCySPAL* achieved strong performance metrics across cross-validation folds (area under the receiver operating characteristic curve (AUROC): 0.964 ± 0.008 (mean ± SD), area under the precision-recall curve (AUPRC): 0.922 ± 0.010, precision: 0.949 ± 0.019, recall: 0.748 ± 0.029, F_1 score: 0.836 ± 0.019) and maintained this performance on the held-out test set (AUROC: 0.964, AUPRC: 0.914, precision: 0.923, recall: 0.736) (Figure 2B).

To assess the relative contributions of the different feature types, ablation studies were performed. A model trained using only SaProt embeddings achieved high performance (AUROC: 0.963, AUPRC: 0.906) on the test set, while a model trained using only the structural descriptors also showed predictive power (AUROC: 0.901, AUPRC: 0.803). The combined *TopCySPAL* model modestly outperformed both individual models (AUROC: 0.964, AUPRC: 0.914), particularly improving the AUPRC. This suggests that while embeddings capture the majority of the predictive signal, explicit structural features provide complementary information that refines the predictions, aligning with the observation that interpretable structural descriptors were among the most important features.

It is important to interpret these performance metrics in the context of the inherent uncertainty associated with negative labels in ligandability prediction. Cysteines labeled as negative (primarily based on non-detectability in proteomics assays) may not be truly non-ligandable but simply may have not been observed to interact under the specific experimental conditions tested so far. Standard classification metrics, calculated against this potentially noisy ground truth, may therefore underestimate the model's ability to

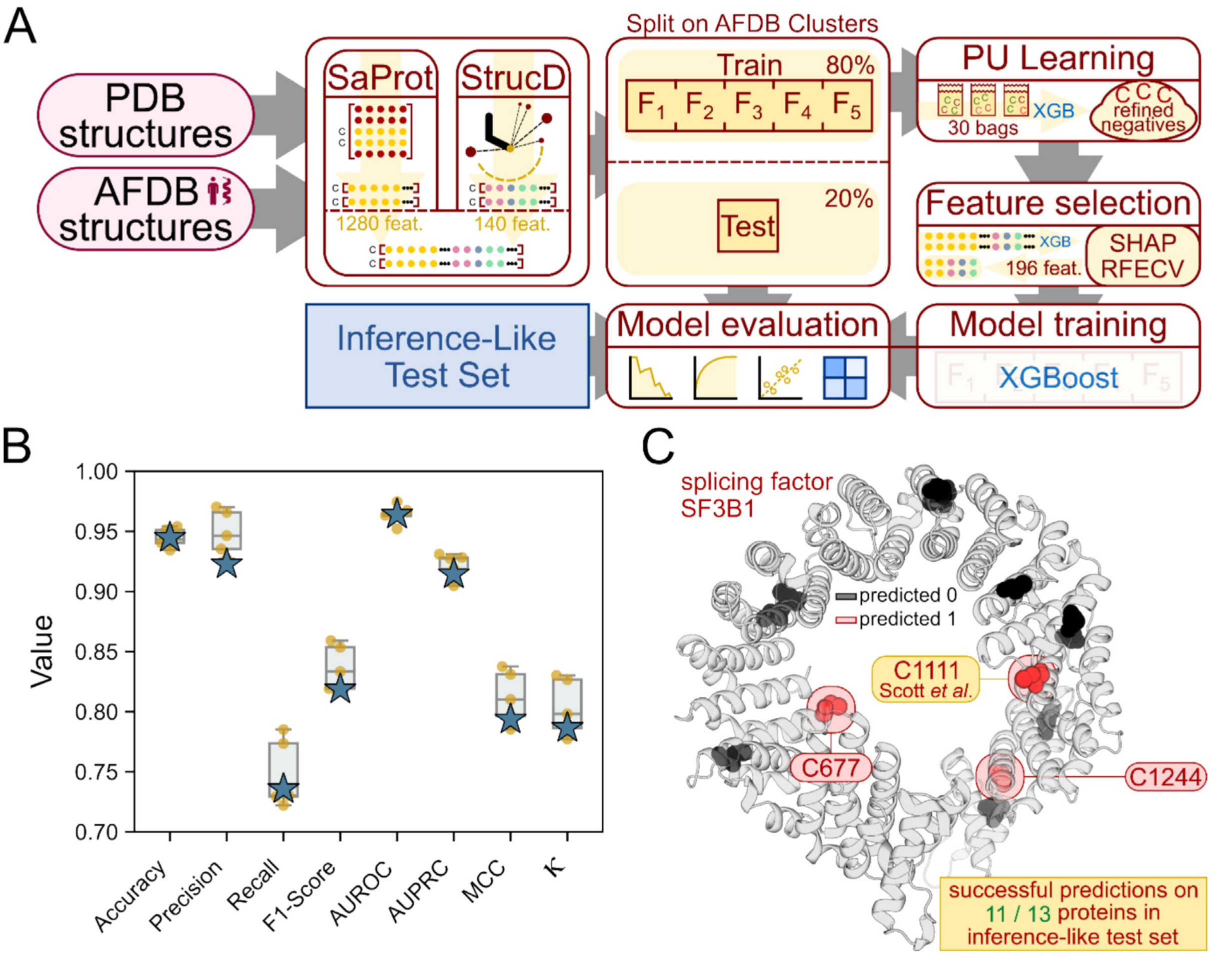


Figure 2. Machine learning pipeline and *TopCySPAL* model performance for cysteine ligandability prediction. (A) PDB and AlphaFold Database (AFDB) structures were processed to generate two types of cysteine features: SaProt embeddings (1,280 features) and structural descriptors (140 features). The term "StrucD" refers to the computational workflow generating the structural descriptors as detailed in SI section 1.2.1.1. The dataset was split into training (80%) and test (20%) sets using FoldSeek-based AFDB clusters [39] to prevent homology-based data leakage. Training data underwent Positive-Unlabeled (PU) learning with 30 independent XGBoost classifier bags to refine negative labels. Feature selection using SHAP-based RFECV led to 196 optimal features. The final XGBoost model was evaluated on the held-out test set and an independent inference-like test set. (B) Distribution of model performance metrics across 5-fold cross-validation (box plots), with individual fold values shown as yellow dots and held-out test set performance as blue stars. AUROC: area under the ROC curve, AUPRC: area under the precision-recall curve, MCC: Matthews correlation coefficient, κ: Cohen's kappa. (C) Example prediction on an AFDB structure of splicing factor SF3B1 (O75533), one of 13 proteins from the inference-like test set. Cysteines are colored according to red: predicted ligandable "1", black: predicted non-ligandable "0". The ABPP-ligandable cysteine C1111 [45,46] was correctly identified, with additional cysteines (C677 and C1244) predicted as potentially ligandable. The box at the bottom refers to the overall result on the inference-like test set.

distinguish truly inert sites from those with future ligandable potential. Our use of PU learning aimed to mitigate this by focusing training on higher-confidence negatives. Ultimately, the model's goal is to generalize beyond current observations and identify potentially novel ligandable sites, a capability supported by its strong performance on the independent inference-like test set (Table SI-1).

Notably, *TopCySPAL* successfully identified 11 out of 13 ABPP-ligandable cysteines in a second inference-like test set derived from literature covering 13 proteins from highly relevant families (e.g., E3 ligases, transcription factors, splicing factors) (see Table SI-1, Figure 2C). These cysteines exhibited proteomics evidence of covalent engagement that was unknown to the model and had been independently validated by corresponding alanine- or serine-mutants, the measurement of binding characteristics to the identified target, or other complementary experiments. For the shown example protein

SF3B1 (O75533), the model correctly identified C1111 [45,46], confirming *TopCySPAL*'s ability to narrow down true positives in noisy non-targeted chemoproteomics approaches, thereby aiding decision-making in covalent drug discovery. In addition, C677 and C1244 were predicted to be ligandable, suggesting their potential as covalent drug targets and warranting further investigation. The false-negative cases involved zinc-binding cysteines in RNF126 [47] and RNF4 [48], highlighting both the model's ability to recognize the negatively labeled metal-binding signature and potential limitations in the overall classification in unusual cases where typical metal-binding cysteines can engage in covalent interactions with TCIs.

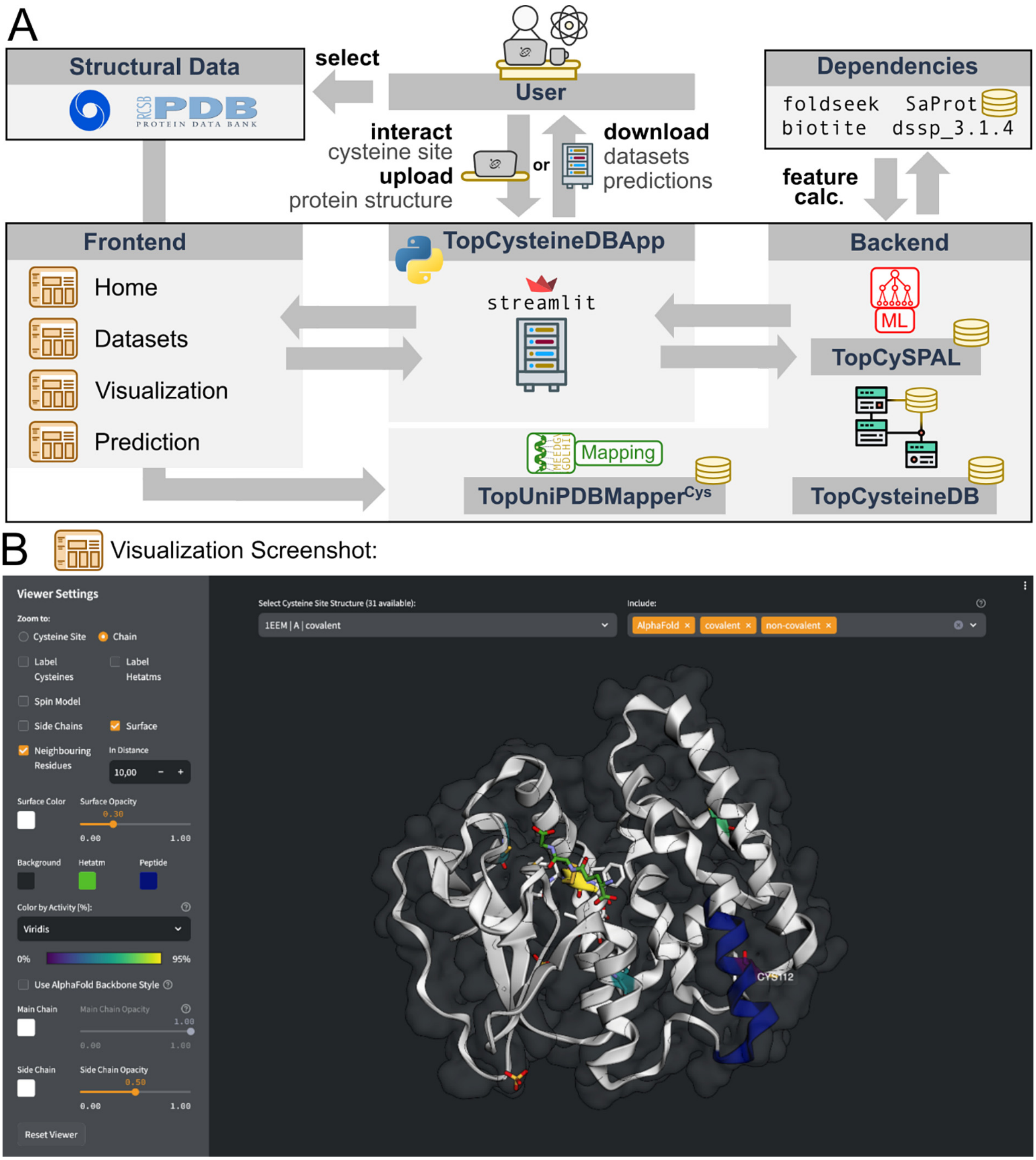
Analysis of feature importance post training revealed that predictions rely on a combination of structural descriptors that match chemical intuition and specific embedding dimensions. The most influential features included side-chain solvent accessibility, secondary structure assignment at the cysteine site, and the atomic density within 6 Å, and the key embedding dimensions are 805, 732, 1025, 1266, 638, 595, and 1083. This combination enables the model to capture complex patterns associated with cysteine ligandability while maintaining physical plausibility, as the highest-ranking structural descriptors align with established chemical and structural principles of cysteine reactivity. The balanced importance of both interpretable structural descriptors and learned embedding features provides confidence in the model's decision-making process, demonstrating the value of integrating both local structural properties and broader sequence-based patterns.

Web interface provides access to integrated cysteine data and predicted ligandability

TopCysteineDB is freely accessible via a web interface generated by the streamlit-based Python package *TopCysteineDBApp* (https://topcysteinedb.hhu.de/), designed to facilitate the exploration of cysteine sites across the human proteome (Figure 3A). The implemented backend enables efficient SQL queries against *TopCysteineDB*, the retrieval of selected cysteine site structures from the PDB or AlphaFoldDB, reliable sequence-structure mapping using *TopUniPDBMapper* (see SI Section 1.2), and ligandability predictions using *TopCySPAL* (features are calculated using Foldseek, SaProt, Biotite, and DSSP; see SI Section 1.2). The application's server can be run locally or operated on a remote machine and permits user interaction with a frontend that consists of four pages. Besides the initial "Home" page, users can browse, filter, and download protein-specific datasets of structural and proteomics data via the "Datasets" page. Moreover, the "Visualization" page uniquely combines visualization of protein structures with the available proteomics data, allowing researchers to examine the classified cysteine site structures and simultaneously assess corresponding reactivity information in its three-dimensional context (Figure 3B). Utilizing the cross-referenced nature of the available residue-wise data in combination with the flexible py3Dmol visualization capabilities [49], a highly customizable viewer has been created that allows for a color-coded mapping of ligandability information onto a protein structure of choice. The style settings of chains, surfaces, cysteine-, neighboring- and hetatm-residues can be adapted. Protein sequence parts corresponding to peptides detected in the proteomics experiments can be highlighted. For a selected cysteine site, a comprehensive overview of the available proteomics data across all quantified ligands is provided. Visualized protein structures and ligands can be downloaded in common file formats. Finally, ligandability predictions by *TopCySPAL* on both existing and uploaded PDB and AlphaFold structures can be performed on the "Prediction" page.

Conclusion

TopCysteineDB represents a significant advancement in understanding cysteine ligandability by bridging structural biology and chemoproteomics data. Through the combination of our *TopCovPDB* classification pipeline with extensive manual curation we have created a high-quality dataset that provides unprecedented granularity in distinguishing different types of covalent modifications while maintaining broad coverage of the human cysteinome. The close integration of diverse structural information and proteomics data combined with state-of-the-art embeddings and physics-based descriptors enabled the development of the machine learning model *TopCySPAL*, which shows strong predictive performance. The model's transferable nature was demonstrated by the successful validation on independently confirmed ABPP-ligandable cysteines, showcasing its practical utility for drug discovery applications supporting the identification of potentially novel ligandable cysteine sites. Through its interactive web interface, *TopCysteineDB* makes the wealth of integrated data easily accessible to researchers, offering both comprehensive visualization tools and predictive capabilities. This resource contributes to the field of covalent ligand discovery by providing integrated access to previously dispersed data types, thereby enabling systematic identification and validation of novel cysteine targets across the human proteome. While the current focus has



Viewer Settings

Zoom to:
○ Cysteine Site ● Chain

□ Label Cysteines □ Label Hetatms

□ Spin Model

□ Side Chains ☑ Surface

☑ Neighbouring Residues In Distance 10,00 − +

Surface Color Surface Opacity 0.30 (0.00 – 1.00)

Background Hetatm Peptide

Color by Activity [%]: Viridis

0% 95%

□ Use AlphaFold Backbone Style

Main Chain Main Chain Opacity 1.00 (0.00 – 1.00)

Side Chain Side Chain Opacity 0.50 (0.00 – 1.00)

Reset Viewer

Select Cysteine Site Structure (31 available): 1EEM | A | covalent

Include: AlphaFold × covalent × non-covalent ×

Figure 3. Interface of the *TopCysteineDB* webserver. (**A**) Schematic representation of the *TopCysteineDBApp* that enables user interaction with *TopCysteineDB* and *TopCySPAL*. See text for details. (**B**) Screenshot of the color-coded proteomics activity mapping accessible via the "Visualization" page, exemplified for a user-selected structure of glutathione *S*-transferase omega-1 (PDB ID 1EEM). The viewer can be customized via the sidebar settings. The maximum CR values known for the structurally resolved cysteines 32 (classified as "covalent"; highly reactive, yellow), 90, 112 (not reactive, violet), 192, and 237 (moderately reactive, blue-green) are indicated by colors (see color scale) corresponding to CR values between 1 and ≥4 (expressed as activity between "0%" and "95%"). The covalent "GSH" ligand identified by *TopCovPDB* is highlighted in green. For cysteine 112, the label has been activated by hovering over the residue. The protein region flanking this residue has been identified via a user-provided peptide sequence and is highlighted in dark blue.

been on human proteins due to the available proteomics data, our approach should be extendable to the analysis of microbial or other pathogen targets, opening possibilities for broader therapeutic applications.

Materials & Methods

TopCysteineDB: A comprehensive database of cysteine site structures and cysteine ligandability

TopCysteineDB integrates structural and chemoproteomics data in an SQLite database linking protein structures, cysteine sites, and experimental chemoproteomics measurements. Using our automated pipeline *TopCovPDB*, we systematically classified cysteine sites based on their interaction partners, followed by manual curation. For detailed methodology, see SI Section 1.1.

Prediction of cysteine ligandability

We developed *TopCySPAL* (Cysteine Structure–Proteomics-Augmented Ligandability predictor), combining structural descriptors and protein embeddings with proteomics data. The model achieved strong performance (AUROC: 0.964, AUPRC: 0.914) and successfully identified 12/13 independently validated ABPP-ligandable cysteines. For detailed methodology, see SI Section 1.2.

Interface

TopCysteineDB is accessible through a web interface (https://topcysteinedb.hhu.de) enabling interactive visualization and ligandability prediction. For detailed methodology, see SI Section 1.3.

CRediT authorship contribution statement

Michele Bonus: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Julian Greb:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jaimeen D. Majmudar:** Writing – review & editing, Validation, Investigation. **Markus Boehm:** Writing – review & editing, Validation, Investigation. **Magdalena Korczynska:** Writing – review & editing, Validation, Investigation. **Azadeh Nazemi:** Writing – review & editing, Validation, Investigation. **Alan M. Mathiowetz:** Writing – review & editing, Validation, Supervision, Project administration, Investigation, Conceptualization. **Holger Gohlke:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Investigation, Funding acquisition, Conceptualization.

Funding

This study was supported by funds from Pfizer, Inc., and, in part, by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Grant 270650915 (Research Training Group GRK 2158).

Data availability

The *TopCysteineDB* can be downloaded as SQLite file from the Heinrich Heine University Research Data server (https://researchdata.hhu.de/handle/entry/76).

DECLARATION OF COMPETING INTEREST

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: ‘JDM, MBoe, MK, AN, and AMM are employees of Pfizer Inc. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.’.

Acknowledgements

We thank Christopher Pfleger for his early contributions that helped shape the foundational concepts of *TopCysteineDB.* Computational support and infrastructure was provided by the “Zentrum für Informations- und Medientechnologie” (ZIM) at Heinrich Heine University Düsseldorf. We thank Nicholas Labello for his valuable suggestions regarding software and platform choices. We are grateful for computing time provided by the John von Neumann Institute for Computing (NIC) to HG on the supercomputer JUWELS at Jülich Supercomputing Centre (JSC) (user ID: VSK33, FOUND).

Appendix A. Supplementary material

Supplementary material to this article can be found online at https://doi.org/10.1016/j.jmb.2025.169196.

Received 15 January 2025;
Accepted 5 May 2025;
Available online 8 May 2025

Keywords:
cysteine;
ligandability;

database;
chemoproteomics;
machine learning

† These authors contributed equally to this work.

Abbreviations:

AFDB, **A**lpha**F**old **D**ata**B**ase; **AUPRC**, **A**rea **U**nder the **P**recision-**R**ecall **C**urve; **CNN**, **C**onvolutional **N**eural **N**etwork; **CR**, **C**ompetition **R**atio; **CV**, **C**oefficient of **V**ariation; **isoTOP-ABPP**, **iso**topic **T**andem **O**rthogonal **P**roteolysis-**A**ctivity-**B**ased **P**rotein **P**rofiling; **KRAS**, **K**irsten **RA**t **S**arcoma virus; **PDB**, **P**rotein **D**ata **B**ank; **PU**, **P**ositive-**U**nlabeled; **SASA**, **S**olvent **A**ccessible **S**urface **A**rea; **SD**, **S**tandard **D**eviation; **SVM**, **S**upport **V**ector **M**achine; **TCI**, **T**argeted **C**ovalent **I**nhibitor

References

[1]. McWhirter, C., (2021). Kinetic mechanisms of covalent inhibition. In: Ward, R.A., Grimster, N. (Eds.), *The Design of Covalent-Based Inhibitors*. Academic Press, Cambridge, Massachusetts, pp. 1–31.

[2]. Singh, J., Petter, R.C., Baillie, T.A., Whitty, A., (2011). The resurgence of covalent drugs. *Nature Rev. Drug Discov.* **10**, 307–317. https://doi.org/10.1038/nrd3410.

[3]. Baillie, T.A., (2016). Targeted covalent inhibitors for drug design. *Angew. Chem. Int. Ed.* **55**, 13408–13421. https://doi.org/10.1002/anie.201601091.

[4]. Lonsdale, R., Ward, R.A., (2018). Structure-based design of targeted covalent inhibitors. *Chem. Soc. Rev.* **47**, 3816–3830. https://doi.org/10.1039/c7cs00220c.

[5]. Singh, J., (2022). The ascension of targeted covalent inhibitors. *J. Med. Chem.* **65**, 5886–5901. https://doi.org/10.1021/acs.jmedchem.1c02134.

[6]. Gehringer, M., Laufer, S.A., (2019). Emerging and re-emerging warheads for targeted covalent inhibitors: applications in medicinal chemistry and chemical biology. *J. Med. Chem.* **62**, 5673–5724. https://doi.org/10.1021/acs.jmedchem.8b01153.

[7]. Hillebrand, L., Liang, X.J., Serafim, R.A.M., Gehringer, M., (2024). Emerging and re-emerging warheads for targeted covalent inhibitors: an update. *J. Med. Chem.* **67**, 7668–7758. https://doi.org/10.1021/acs.jmedchem.3c01825.

[8]. Maurais, A.J., Weerapana, E., (2019). Reactive-cysteine profiling for drug discovery. *Curr. Opin. Chem. Biol.* **50**, 29–36. https://doi.org/10.1016/j.cbpa.2019.02.010.

[9]. Lu, X., Smaill, J.B., Patterson, A.V., Ding, K., (2022). Discovery of cysteine-targeting covalent protein kinase inhibitors. *J. Med. Chem.* **65**, 58–83. https://doi.org/10.1021/acs.jmedchem.1c01719.

[10]. Hong, D.S., Fakih, M.G., Strickler, J.H., Desai, J., Durm, G.A., Shapiro, G.I., Falchook, G.S., Price, T.J., Sacher, A., Denlinger, C.S., Bang, Y.-J., Dy, G.K., Krauss, J.C., Kuboki, Y., Kuo, J.C., Coveler, A.L., Park, K., Kim, T.W., Barlesi, F., Munster, P.N., Ramalingam, S.S., Burns, T.F., Meric-Bernstam, F., Henary, H., Ngang, J., Ngarmchamnanrith, G., Kim, J., Houk, B.E., Canon, J., Lipford, J.R., Friberg, G., Lito, P., Govindan, R., Li, B.T., (2020). KRAS(G12C) Inhibition with sotorasib in advanced solid tumors. *N. Engl. J. Med.* **383**, 1207–1217. https://doi.org/10.1056/NEJMoa1917239.

[11]. Niphakis, M.J., Cravatt, B.F., (2024). Ligand discovery by activity-based protein profiling. *Cell Chem. Biol.* **31**, 1636–1651. https://doi.org/10.1016/j.chembiol.2024.08.006.

[12]. Bar-Peled, L., Kemper, E.K., Suciu, R.M., Vinogradova, E.V., Backus, K.M., Horning, B.D., Paul, T.A., Ichu, T.-A., Svensson, R.U., Olucha, J., Chang, M.W., Kok, B.P., Zhu, Z., Ihle, N.T., Dix, M.M., Jiang, P., Hayward, M.M., Saez, E., Shaw, R.J., Cravatt, B.F., (2017). Chemical proteomics identifies druggable vulnerabilities in a genetically defined cancer. *Cell* **171**, 696–709.e623. https://doi.org/10.1016/j.cell.2017.08.051.

[13]. Vinogradova, E.V., Zhang, X., Remillard, D., Lazar, D.C., Suciu, R.M., Wang, Y., Bianco, G., Yamashita, Y., Crowley, V.M., Schafroth, M.A., Yokoyama, M., Konrad, D.B., Lum, K.M., Simon, G.M., Kemper, E.K., Lazear, M. R., Yin, S., Blewett, M.M., Dix, M.M., Nguyen, N., Shokhirev, M.N., Chin, E.N., Lairson, L.L., Melillo, B., Schreiber, S.L., Forli, S., Teijaro, J.R., Cravatt, B.F., (2020). An activity-guided map of electrophile-cysteine interactions in primary human T cells. *Cell* **182**, 1009–1026.e1029. https://doi.org/10.1016/j.cell.2020.07.001.

[14]. Kuljanin, M., Mitchell, D.C., Schweppe, D.K., Gikandi, A. S., Nusinow, D.P., Bulloch, N.J., Vinogradova, E.V., Wilson, D.L., Kool, E.T., Mancias, J.D., Cravatt, B.F., Gygi, S.P., (2021). Reimagining high-throughput profiling of reactive cysteines for cell-based screening of large electrophile libraries. *Nature Biotechnol.* **39**, 630–641. https://doi.org/10.1038/s41587-020-00778-3.

[15]. Takahashi, M., Chong, H.B., Zhang, S., Yang, T.-Y., Lazarov, M.J., Harry, S., Maynard, M., Hilbert, B., White, R.D., Murrey, H.E., Tsou, C.-C., Vordermark, K., Assaad, J., Gohar, M., Dürr, B.R., Richter, M., Patel, H., Kryukov, G., Brooijmans, N., Alghali, A.S.O., Rubio, K., Villanueva, A., Zhang, J., Ge, M., Makram, F., Griesshaber, H., Harrison, D., Koglin, A.-S., Ojeda, S., Karakyriakou, B., Healy, A., Popoola, G., Rachmin, I., Khandelwal, N., Neil, J.R., Tien, P.-C., Chen, N., Hosp, T., van den Ouweland, S., Hara, T., Bussema, L., Dong, R., Shi, L., Rasmussen, M.Q., Domingues, A.C., Lawless, A., Fang, J., Yoda, S., Nguyen, L.P., Reeves, S.M., Wakefield, F.N., Acker, A., Clark, S.E., Dubash, T., Kastanos, J., Oh, E., Fisher, D. E., Maheswaran, S., Haber, D.A., Boland, G.M., Sade-Feldman, M., Jenkins, R.W., Hata, A.N., Bardeesy, N.M., Suvà, M.L., Martin, B.R., Liau, B.B., Ott, C.J., Rivera, M. N., Lawrence, M.S., Bar-Peled, L., (2024). DrugMap: a quantitative pan-cancer analysis of cysteine ligandability. *Cell* **187**, 2536–2556.e2530. https://doi.org/10.1016/j.cell.2024.03.027.

[16]. 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., (2016). Proteome-wide covalent ligand discovery in native biological systems. *Nature* **534**, 570–574. https://doi.org/10.1038/nature18002.

[17]. Boatner, L.M., Palafox, M.F., Schweppe, D.K., Backus, K. M., (2023). CysDB: a human cysteine database based on experimental quantitative chemoproteomics. *Cell Chem. Biol.* **30**, 683–698.e683. https://doi.org/10.1016/j.chembiol.2023.04.004.

[18]. Gao, M., Moumbock, A.F.A., Qaseem, A., Xu, Q., Günther, S., (2022). CovPDB: a high-resolution coverage of the covalent protein-ligand interactome. *Nucleic Acids Res.* **50**, D445–D450. https://doi.org/10.1093/nar/gkab868.

[19]. Du, H., Gao, J., Weng, G., Ding, J., Chai, X., Pang, J., Kang, Y., Li, D., Cao, D., Hou, T., (2021). CovalentInDB: a comprehensive database facilitating the discovery of covalent inhibitors. *Nucleic Acids Res.* **49**, D1122–D1129. https://doi.org/10.1093/nar/gkaa876.

[20]. Du, H., Zhang, X., Wu, Z., Zhang, O., Gu, S., Wang, M., Zhu, F., Li, D., Hou, T., Pan, P., (2024). CovalentInDB 2.0: an updated comprehensive database for structure-based and ligand-based covalent inhibitor design and screening. *Nucleic Acids Res.*. https://doi.org/10.1093/nar/gkae946.

[21]. Wu, S., Luo Howard, H., Wang, H., Zhao, W., Hu, Q., Yang, Y., (2016). Cysteinome: The first comprehensive database for proteins with targetable cysteine and their covalent inhibitors. *Biochem. Biophys. Res. Commun.* **478**, 1268–1273. https://doi.org/10.1016/j.bbrc.2016.08.109.

[22]. Du, J., Yan, X., Liu, Z., Cui, L., Ding, P., Tan, X., Li, X., Zhou, H., Gu, Q., Xu, J., (2017). cBinderDB: a covalent binding agent database. *Bioinformatics* **33**, 1258–1260. https://doi.org/10.1093/bioinformatics/btw801.

[23]. Guo, X.-K., Zhang, Y., (2022). CovBinderInPDB: a structure-based covalent binder database. *J. Chem. Inf. Model.* **62**, 6057–6068. https://doi.org/10.1021/acs.jcim.2c01216.

[24]. Awoonor-Williams, E., Rowley, C.N., (2016). Evaluation of methods for the calculation of the pKa of cysteine residues in proteins. *J. Chem. Theory Comput.* **12**, 4662–4673. https://doi.org/10.1021/acs.jctc.6b00631.

[25]. Harris, R.C., Liu, R., Shen, J., (2020). Predicting reactive cysteines with implicit-solvent-based continuous constant pH molecular dynamics in amber. *J. Chem. Theory Comput.* **16**, 3689–3698. https://doi.org/10.1021/acs.jctc.0c00258.

[26]. Soylu, İ., Marino, S.M., (2016). Cy-preds: an algorithm and a web service for the analysis and prediction of cysteine reactivity. *Proteins* **84**, 278–291. https://doi.org/10.1002/prot.24978.

[27]. Wang, H., Chen, X., Li, C., Liu, Y., Yang, F., Wang, C., (2018). Sequence-based prediction of cysteine reactivity using machine learning. *Biochemistry* **57**, 451–460. https://doi.org/10.1021/acs.biochem.7b00897.

[28]. Zhang, W., Pei, J., Lai, L., (2017). Statistical analysis and prediction of covalent ligand targeted cysteine residues. *J. Chem. Inf. Model.* **57**, 1453–1460. https://doi.org/10.1021/acs.jcim.7b00163.

[29]. Gao, M., Günther, S., (2023). HyperCys: a structure- and sequence-based predictor of hyper-reactive druggable cysteines. *Int. J. Mol. Sci.* **24** https://doi.org/10.3390/ijms24065960.

[30]. Du, H., Jiang, D., Gao, J., Zhang, X., Jiang, L., Zeng, Y., Wu, Z., Shen, C., Xu, L., Cao, D., Hou, T., Pan, P., (2022). Proteome-wide profiling of the covalent-druggable cysteines with a structure-based deep graph learning network. *Research* **2022**, 9873564. https://doi.org/10.34133/2022/9873564.

[31]. Liu, R., Clayton, J., Shen, M., Bhatnagar, S., Shen, J., (2024). Machine learning models to interrogate proteome-wide covalent ligandabilities directed at cysteines. *JACS Au* **4**, 1374–1384. https://doi.org/10.1021/jacsau.3c00749.

[32]. Cao, J., Xu, Y., (2024). Predicting cysteine reactivity changes upon phosphorylation using XGBoost. *FEBS Open Bio* **14**, 51–62. https://doi.org/10.1002/2211-5463.13737.

[33]. Reimer, B.M., Awoonor-Williams, E., Golosov, A.A., Hornak, V., (2025). CovCysPredictor: predicting selective covalently modifiable cysteines using protein structure and interpretable machine learning. *J. Chem. Inf. Model.* **65**, 544–553. https://doi.org/10.1021/acs.jcim.4c01281.

[34]. Boatner, L., Eberhardt, J., Shikwana, F., Holcomb, M., Lee, P., Houk, K., Forli, S., Backus, K., (2025). CIAA: integrated proteomics and structural modeling for understanding cysteine reactivity with iodoacetamide alkyne (version 1). *ChemRxiv.* https://doi.org/10.26434/chemrxiv-2025-tm8ch.

[35]. Su, J., Han, C., Zhou, Y., Shan, J., Zhou, X., Yuan, F., (2024). SaProt: protein language modeling with structure-aware vocabulary. *bioRxiv*2023.2010.2001.560349. https://doi.org/10.1101/2023.10.01.560349.

[36]. Velankar, S., Dana, J.M., Jacobsen, J., van Ginkel, G., Gane, P.J., Luo, J., Oldfield, T.J., O'Donovan, C., Martin, M.-J., Kleywegt, G.J., (2013). SIFTS: structure integration with function, taxonomy and sequences resource. *Nucleic Acids Res.* **41**, D483–D489. https://doi.org/10.1093/nar/gks1258.

[37]. Dana, J.M., Gutmanas, A., Tyagi, N., Qi, G., O'Donovan, C., Martin, M., Velankar, S., (2019). SIFTS: updated structure integration with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Res.* **47**, D482–D489. https://doi.org/10.1093/nar/gky1114.

[38]. Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M., Tsenkov, M., Nair, S., Mirdita, M., Yeo, J., Kovalevskiy, O., Tunyasuvunakool, K., Laydon, A., Žídek, A., Tomlinson, H., Hariharan, D., Abrahamson, J., Green, T., Jumper, J., Birney, E., Steinegger, M., Hassabis, D., Velankar, S., (2024). AlphaFold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.* **52**, D368–D375. https://doi.org/10.1093/nar/gkad1011.

[39]. Barrio-Hernandez, I., Yeo, J., Jänes, J., Mirdita, M., Gilchrist, C.L.M., Wein, T., Varadi, M., Velankar, S., Beltrao, P., Steinegger, M., (2023). Clustering predicted structures at the scale of the known protein universe. *Nature* **622**, 637–645. https://doi.org/10.1038/s41586-023-06510-w.

[40]. Bekker, J., Davis, J., (2020). Learning from positive and unlabeled data: a survey. *Mach. Learn.* **109**, 719–760. https://doi.org/10.1007/s10994-020-05877-5.

[41]. Kunzmann, P., Hamacher, K., (2018). Biotite: a unifying open source computational biology framework in Python. *BMC Bioinf.* **19**, 346. https://doi.org/10.1186/s12859-018-2367-z.

[42]. Kunzmann, P., Müller, T.D., Greil, M., Krumbach, J.H., Anter, J.M., Bauer, D., Islam, F., Hamacher, K., (2023). Biotite: new tools for a versatile Python bioinformatics library. *BMC Bioinf.* **24**, 236. https://doi.org/10.1186/s12859-023-05345-6.

[43]. Kabsch, W., Sander, C., (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637. https://doi.org/10.1002/bip.360221211.

[44]. Lundberg, S.M., Lee, S.-I., (2017). A unified approach to interpreting model predictions. *Adv. Neur. In.* **30**

[45]. Lazear, M.R., Remsberg, J.R., Jaeger, M.G., Rothamel, K., Her, H.-L., DeMeester, K.E., Njomen, E., Hogg, S.J., Rahman, J., Whitby, L.R., Won, S.J., Schafroth, M.A., Ogasawara, D., Yokoyama, M., Lindsey, G.L., Li, H., Germain, J., Barbas, S., Vaughan, J., Hanigan, T.W., Vartabedian, V.F., Reinhardt, C.J., Dix, M.M., Koo, S.J., Heo, I., Teijaro, J.R., Simon, G.M., Ghosh, B., Abdel-Wahab, O., Ahn, K., Saghatelian, A., Melillo, B., Schreiber, S.L., Yeo, G.W., Cravatt, B.F., (2023). Proteomic discovery of chemical probes that perturb protein complexes in human cells. *Mol. Cell* **83**, 1725–1742.e1712. https://doi.org/10.1016/j.molcel.2023.03.026.

[46]. Scott, K.A., Kojima, H., Ropek, N., Warren, C.D., Zhang, T.L., Hogg, S.J., Webster, C., Zhang, X., Rahman, J., Melillo, B., Cravatt, B.F., Lyu, J., Abdel-Wahab, O., Vinogradova, E.V., (2023). Covalent targeting of splicing in T cells. *bioRxiv*2023.2012.2018.572199. https://doi.org/10.1101/2023.12.18.572199.

[47]. Toriki, E.S., Papatzimas, J.W., Nishikawa, K., Dovala, D., Frank, A.O., Hesse, M.J., Dankova, D., Song, J.-G., Bruce-Smythe, M., Struble, H., Garcia, F.J., Brittain, S.M., Kile, A.C., McGregor, L.M., McKenna, J.M., Tallarico, J. A., Schirle, M., Nomura, D.K., (2023). Rational chemical design of molecular glue degraders. *ACS Cent. Sci.* **9**, 915–926. https://doi.org/10.1021/acscentsci.2c01317.

[48]. Ward, C.C., Kleinman, J.I., Brittain, S.M., Lee, P.S., Chung, C.Y.S., Kim, K., Petri, Y., Thomas, J.R., Tallarico, J.A., McKenna, J.M., Schirle, M., Nomura, D.K., (2019). Covalent ligand screening uncovers a RNF4 E3 ligase recruiter for targeted protein degradation applications. *ACS Chem. Biol.* **14**, 2430–2440. https://doi.org/10.1021/acschembio.8b01083.

[49]. Rego, N., Koes, D., (2015). 3Dmol.js: molecular visualization with WebGL. *Bioinformatics* **31**, 1322–1324. https://doi.org/10.1093/bioinformatics/btu829.