Illuminating the Druggable Proteome with an AI Protein Profiling Platform

Guy W. Dayhoff II,[†] Daniel Kortzak,[†,‡] Ruibin Liu,[†] Mingzhe Shen,[†] and Jana Shen[*,†]

† *Department of Pharmaceutical Sciences, University of Maryland School of Pharmacy, Baltimore, MD 21201, U.S.A.*
‡ *Joint first author*

E-mail: jana.shen@rx.umaryland.edu

Abstract

Creating a proteomewide atlas of reversible and covalent protein-ligand binding sites would transform our understanding of protein functions and accelerate therapeutic discovery; however, current approaches face significant challenges. Chemical proteomics is constrained by limited proteome coverage and data heterogeneity, while existing machine learning models have limited predictive power due to structural dependencies and inadequate handling of heterogeneous experimental labels. Here we developed AiPP, an AI protein profiling platform that provides comprehensive annotations of ligand interaction sites directly from sequence. AiPP is powered by the evolutionary scale protein large language models (ESM3 and ESMC) and two new databases, LigCysABPP (cysteine ligandability from activity-based protein profiling or ABPP) and LigBind3D (reversible binding sites in co-crystal structures). LigCysABPP comprises >700,000 cysteine-site records from 15 ABPP studies covering >10,000 human proteins. We developed an evolutionary representation-based clustering framework followed by consensus analysis to reconcile and augment heterogeneous experimental annotations. Two complementary iterative data expansion protocols were implemented to enhance model performance and generalization. AiPP recovers 80% (Top-1) of all cysteine liganding events from the Protein Data Bank with AUPRC of 84% and AUROC of 89%. Furthermore, it recapitulates the consistently and heterogeneously liganded cysteines determined by a recent ABPP study using over 400 cancer cell lines. AiPP provides a starting point for creating a proteome-wide atlas of protein-ligand interactions and systematic discovery of druggable sites. The integrative approach that leverages evolutionary information to harmonize large-scale proteomic data can be broadly applied to protein research.

Introduction

The human proteome comprises ∼20,300 distinct genes that encode over 200,000 proteoforms[1] (proteins and their variants arising from splicing, post-translational modifications, single-nucleotide polymorphisms and mutations); however, fewer than 900 proteins have been therapeutically targeted by the FDA-approved drugs (`https://go.drugbank.com/`).[2] This gap underscores both the urgent need and vast potential to expand the druggable proteome. First developed as a chemical proteomic approach that utilizes small-molecule covalent probes (known as the activity-based probes) to report on enzyme

activities in complex biological systems,[3,4] activity-based protein profiling (ABPP) has evolved into a powerful and versatile technology. By combining activity- or reactivity-based probes with mass spectrometry (MS) based quantitative proteomics, ABPP has enabled several major advances,[5,6] including the discovery of reactive cysteines[7] and characterization of small-molecule–protein interactions (ligandability) in cells,[8] as well as the development of covalent inhibitors.[9] In particular, the isotopic tandem orthogonal proteolysis (isoTOP) ABPP experiments utilizing competition ratios (CRs) between covalent probes and fragments first introduced by Cravatt and Weerapana[7,10,11] have led to quantitative assessment of cysteine ligandability across over 10,000 proteins using different human cell lines[8,12–24] Recently, the ABPP technology has been further developed for high-throughput screening through minimizing the amount of starting protein materials and instrument time[15] as well as using label-free quantification.[18,25]

Despite the enormous potential of ABPP and rapid progress made in recent years, several challenges remain, hindering efforts of employing ABPP to create a comprehensive, accurate map of cysteine-directed ligandable sites across the proteome for covalent drug design. One major challenge is limited proteome coverage due to the dependence on protein abundance, probe chemistry, and cell type (e.g., membrane-bound proteins are underrepresented in lysates due to insolubility).[5,25–30] A second major challenge is the variability in detected proteins, cysteines and their assessed ligandability across the published ABPP datasets (see refs[25,30–32] and later discussions). This variability may at least in part be attributed to the diverse cellular contexts[25,31] and to specific experimental conditions,[30] including probe chemistry and concentration, fragment concentration, treatment duration, cell type (e.g., cell lysates vs. live cells), cellular state (e.g., mitotic vs. asynchronous conditions; protein mutations; post-translational modifications; redox state),[33] cell line (e.g., different cancers[31]), specific electrophilic fragments used,[25] and quantification protocols (e.g., MS protocol and CR threshold for defining a liganding event). An additional challenge arises from the inherent limitations of shotgun proteomics,[34] which prevents the unambiguous identification of a subset of ligandable cysteines.[18,25] To address these limitations, recent studies implemented label-free and data-independent acquisition protocols,[18,25] demonstrating improved proteome coverage depth and enhanced data completeness compared to traditional approaches.

In contrast to proteomics approaches which offer broad coverage across the proteome but with significant data variability, X-ray crystallography delivers high-resolution structural information on protein-ligand interactions for a limited number of complexes. Drawing from the cysteine-liganded co-crystal structures deposited in the Protein Data Bank (PDB), several databases (e.g., CovPDB,[35] CovBinderInPDB,[36] and LigCys3D[37]) have been developed, which annotate nearly 800 proteins with chemically modified cysteines. Based on these databases and experimental or AlphaFold[38] structure models, machine learning (ML) models, including support vector models (SVMs),[39] boosted decision trees,[37,40] graph neural networks (GNNs),[41] and convolutional neural networks (CNNs),[37] have been developed for structure-based prediction of ligandable cysteines, achieving recall and precision as high as 95%. However, these models are not yet deployable for proteome-wide mapping of ligandable cysteine sites due to three critical limitations: (i) the extremely small training dataset and lack of protein diversity; (ii) the absence of reliable controls for unligandable cysteines; and (iii) the requirement for protein structures, which often cannot be met even with state-of-the-art structure prediction tools such as AlphaFold3.[42]

Recently, the ABPP-derived databases (CysDB[43] and TopCysteineDB[32]) have been developed and used to train structure-based ML models for predicting cysteine reactivity or ligandability.[32,44] Based on CysDB[43] and other ABPP data (∼800 reactive cys-

teines) and associated crystal structures, Forli, Backus, and coworkers developed the random forest model CIAA to predict reactive cysteines with an accuracy of 68% in external validation.[44] Using TopCysteineDB, which combines PDB- (nearly 800 liganded cysteines) with ABPP-derived cysteine ligandability data (over 9,000 ligandable cysteines compiled from three publications[13,15,31]) and experimental or AlphaFold structures, Gohlke and coworkers[32] developed the XGBoost model TopCySPAL to predict cysteine ligandability. A unique feature of TopCySPAL is that it is trained with negative cysteines defined as those undetected in liganded proteins quantified by ABPP.[32] In the 20% hold-out test, TopCySPAL achieved AUROC and AUPRC of 96.4% and 91.4%, respectively. Another structure-based ML model was the CNN[31] developed using the DrugMap database (an ABPP dataset obtained with 400 cancer cell lines), which achieved AUROC of 69.3% in external evaluation.

Despite recent advances, existing ML models for predicting cysteine ligandability remain severely constrained in both performance and applicability. First, current approaches for developing ML models inadequately leverage ABPP datasets and lack rigorous frameworks for assignment of ligandability labels. These models fail to reconcile the inherent variability in experimental labels, leading to inconsistent training data and unreliable predictions. Second, this reconciliation challenge is compounded by the absence of standardized or rigorous evaluation metrics and benchmarking protocols, making it difficult to rigorously assess model performance or compare different approaches in a meaningful way. Finally, and perhaps most critically, existing models require high-quality protein structures as input, rendering them inapplicable to substantial portions of the proteome that remain structurally unresolved even with state-of-the-art structure prediction tools such as AlphaFold3,[42] creating a significant coverage gap in ligandability assessment.

To overcome the ML challenges and unlock the full potential of ABPP, we developed a sequence-based AI Protein Profiling (AiPP) platform for annotating reversible ligand binding (LigBind) and (cysteine-directed) covalently ligandable residues (LigCys) across the proteome. AiPP leverages the state-of-the-art protein large language models (pLLM) ESM Cambrian (ESMC)[45] as the foundation model. We curated a comprehensive ABPP database LigCysABPP comprising 15 independent datasets published since 2016. We developed a representation-based clustering strategy to reconcile and augment the ligandability labels. A similar strategy is applied to refine the labels in LigBind3D. We further developed two iterative learning protocols to systematically expand the ML training data, achieving enhanced model accuracy and generalization. Beyond predicting ligandable sites, AiPP incorporates RIDA, a disorder annotation platform that identifies short disordered molecular recognition features (MoRFs) with a propensity for binding-induced folding Cysteine-containing MoRFs are particularly interesting, as they may serve as anchor points for protein-protein interaction (PPI) drug design. AiPP provides a unified platform, paving the way for creating proteome-wide annotation of druggable pockets and their adjacent covalently ligandable residues.

Results and Discussion

Curation of an ML-ready cysteine ligandability database LigCysABPP. From 15 independent ABPP datasets (also referred to as sources) published between 2016 and 2025,[8,12–25] we extracted 683,192 peptide-derived records of cysteine ligandability for 58,704 unique cysteine residues across 14,417 proteins in human proteomes. Following data clean-up steps, including protein sequence reconstruction, record validation, and consolidation of UniProt IDs (UIDs) referring to identical protein sequences, the raw entries were converted to 669,908 structured, site-specific cysteine ligandability records for 53,867 cysteine residues across 11,017 pro-

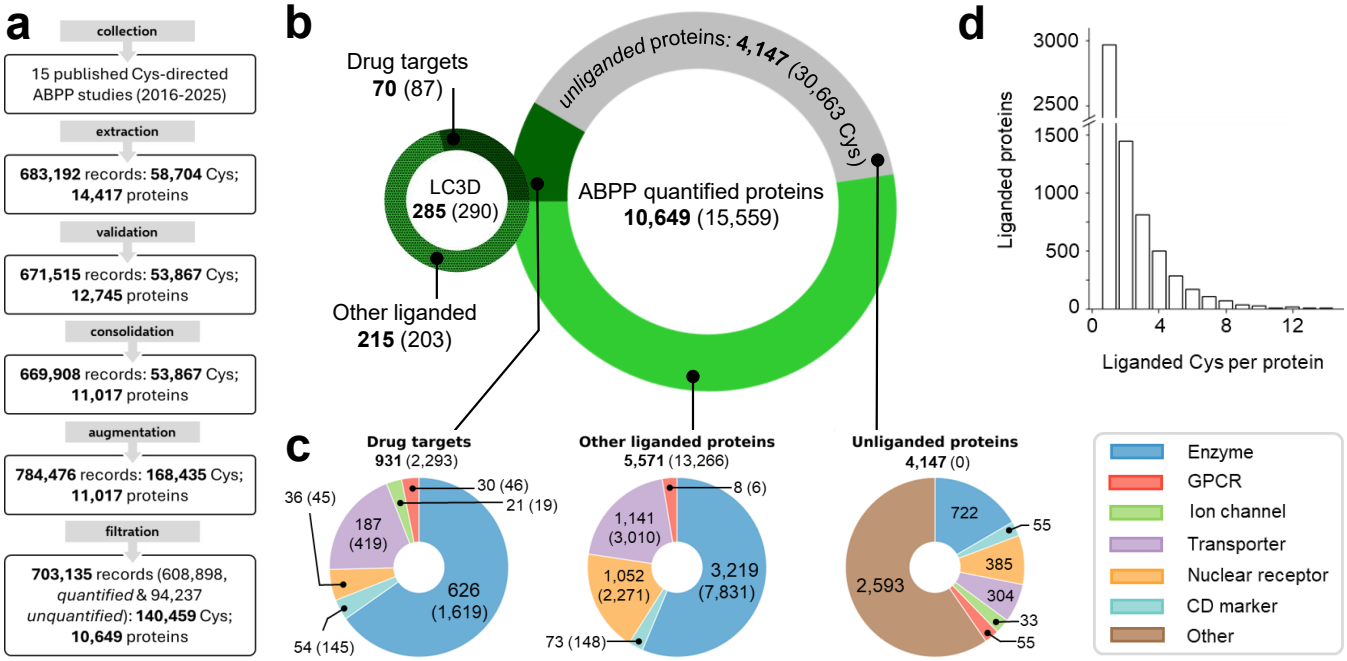

Figure 1: **Overview of the ABPP-quantified proteins and cysteines in LigCysABPP database. a.** The LigCysABPP database is manually curated from a collection of 15 published cysteine-directed ABPP studies.[8,12–25] Peptide-derived records are extracted from each publication, validated against UniProt, consolidated across identical sequences and subsequences, augmented with unquantified (unseen) cysteine-sites on otherwise quantified proteins, and filtered for compatibility with downstream processing steps. **b.** LigCysABPP contains the cysteine ligandability records of 10,649 ABPP-quantified proteins, of which 6,502 are liganded and 4,147 are unliganded. A liganded cysteine is defined as having at least 1 pos ABPP record while a liganded protein is defined as having at least 1 liganded cysteine. Among the liganded proteins, 931 are known drug targets (with 2,293 liganded cysteines), while the rest of 5,571 other liganded proteins contain 13,266 liganded cysteines. The LC3D database contains 285 unique proteins (290 cysteines liganded in the co-crystal structures), among which only 70 proteins (87 drug targets) are also identified as liganded by ABPP. **c.** Functional classes of drug targets, other liganded proteins, and unliganded proteins according to The Human Protein Atlas classifications[46] (`https://www.proteinatlas.org/`). **d.** Histogram of the number of ABPP liganded cysteines per protein shows that most liganded proteins contain 1-3 liganded cysteines.

teins. Each record contains the ligandability label (pos for liganded and neg for unliganded) given by the study authors for a specific cysteine site in a protein under a set of experimental conditions and cellular context. To fully characterize the cysteines in ABPP-quantified proteins, we added unquantified (provisionally neg) records for cysteines belonging to the ABPP-quantified proteins but undetected by the probes. To accommodate the ESM-based ML, protein sequences shorter than 30 or longer than 2046 amino acids were removed. This procedure results in 703,135 total curated records describing ligandability (pos, neg, or unquantified) of 140,459 unique cysteine sites across 10,649 distinct proteins (Fig. 1a). Detailed extraction, validation, and consolidation logs are provided in Supplemental Data (SD6–SD11).

Functional analysis of ABPP-quantified proteins. We define a liganded cysteine as one with at least one positive ABPP record. A supporting source for a liganded cysteine is one that contains at least one pos record for that cysteine. Among the 10,649 ABPP-quantified proteins, 6,502 are liganded (containing at least one liganded cysteine) and 4,147 are unliganded (Fig. 1b). Among the liganded proteins, 931 (with 2,293 pos cysteines) are known drug targets, while the rest of 5,571 proteins (with 13,266 pos cysteines) represent unexplored potential therapeutic opportunities (Fig. 1b). Among the liganded drug targets and other proteins, enzymes constitute the largest functional

family, followed by transporters, nuclear receptors, and others (Fig. 1c). It is noteworthy that transporters and nuclear receptors are substantially more prevalent among other ligandable proteins, each representing over one-third the abundance of enzymes (Fig. 1c). Interestingly, GPCRs are sparsely represented, likely because most cysteine residues are not positioned near ligand-binding pockets. It is intriguing that the majority of unliganded proteins (2,593 out of 4,147) do not belong to any functional class identified by The Human Protein Atlas[46] (`https://www.proteinatlas.org/`), suggesting that a significant fraction of the human proteome remains to be functionally characterized.

The overlap between the ABPP-liganded proteins and those evidenced by co-crystal structures in PDB is extremely small, only 60 proteins (42 are known drug targets, Fig. 1b). This limited overlap may be attributed to the historical focus of structural studies on bacterial and other non-human proteins, which are overrepresented in the PDB. The limited overlap also demonstrates the potential of ABPP to greatly expand the known ligandable proteome. The vast ABPP data also allows us to ask a pertinent question: how many ligandable cysteines are in a single protein? Interestingly, while the majority of proteins contain only one ligandable cysteine, a sizable number of proteins harbor 2–4 ligandable cysteines (Fig. 1d), suggesting that cysteine-directed probes or inhibitors may be designed to target allosteric and other ligandable pockets.

Analysis of ABPP sources and records to assess consensus in cysteine ligandability. In order to devise an ML labeling scheme based on LigCysABPP records, we first analyzed the consensus among 15 sources for liganded cysteines, which are grouped by the number of sources that quantified them (n =1,...,12). For instance, 2,996 cysteines were quantified by a single source, while 2,644 and 2,154 were quantified by 2 and 3 sources, respectively. We then examined the consensus levels by determining how many liganded cysteines are supported by m number of sources (m =1,...,8). This analysis highlights the challenge of reconciling source-specific labels, as considerable variability exists. For example, among cysteines quantified by 2 sources, only 310 (11%) are supported by both sources, while 2,334 (89%) showed conflicting results – pos in one source and neg in the other. Among cysteines quantified by 3 sources, only 82 (4%) are unanimously supported, while 449 (21%) are pos in 2 sources and 1,623 (75%) are pos in only 1 source. This pattern of decreasing consensus becomes more pronounced as the number of sources increases (Fig. 2a and Supplemental Fig. S1).

In the above analysis, a supporting source can have any number of pos records. To further dissect the ABPP noise, we also examined the number of supporting records for liganded cysteines (Fig. 2b). For liganded cysteines supported by 1 source, 7,787 have only 1 pos record, 1,374 have 2 pos records, and 816 have 3–4 pos records. For liganded cysteines supported by 2 sources, 814 have 3–4 positive records. As the number of supporting sources increases, the number of liganded cysteines with positive records decreases – with 5 sources, the number of liganded cysteines with at least 5 pos records is below 30. The steep decline in well-supported liganded cysteines suggests that requiring agreement among more than 4 sources may be too stringent for practical ML applications, as it would severely limit the training data size.

The significant variation in cysteine ligandability likely reflects differences in experimental conditions (e.g., probe and fragment concentrations) and cellular context (e.g., cell line-specific factors) as shown in SI Data Table S1. The influence of cellular context on cysteine ligandability was recently[31] demonstrated by an ABPP study of over 400 cancer proteomes, which identified nearly 400 cysteines exhibiting varying degrees of engagement by an identical fragment ("heterogeneous ligandable"). While cellular con-

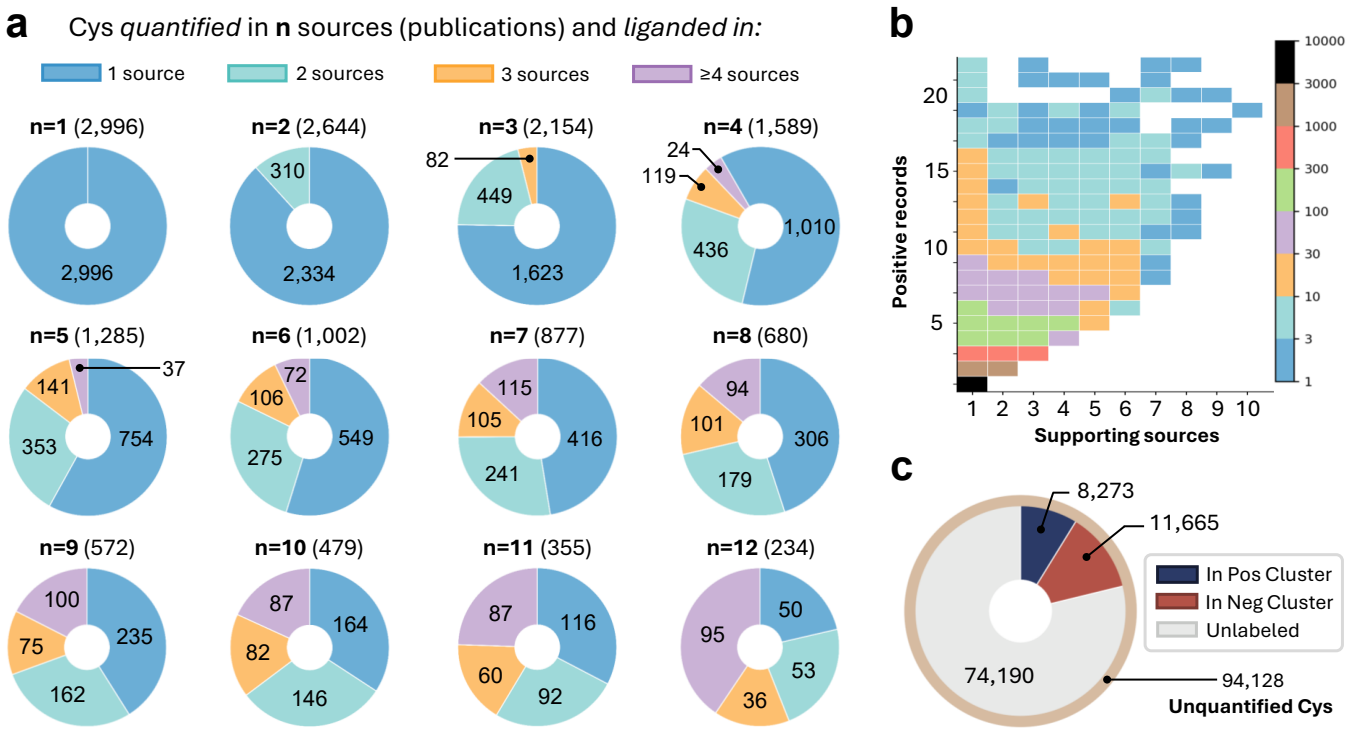

Figure 2: **Analysis of cysteine ligandability across ABPP sources and records. a.** Each pie chart displays the number of liganded cysteines quantified by n sources. A liganded cysteine is defined as one that has at least one pos ABPP record. Each pie segment represents the number of cysteines supported by 1 (blue), 2 (green), 3 (yellow), and $\geq$4 (purple) sources. **b.** The number of liganded cysteines with the number of supporting sources and positive records. A supporting source for a specific cysteine is defined as one that contains at least one pos record for that cysteine. A total of 7,787 cysteines are labeled pos by only 1 record (black). No liganded cysteine is supported by $>$10 sources. Liganded cysteines with $>$25 records are negligible and therefore excluded from the plot. **c.** Following representation-based clustering, 8,273 unquantified cysteines are labeled pos, 11,665 labeled neg, while 74,190 cysteines remain unlabeled.

texts will be addressed in our future studies, the present work focuses on cysteines that demonstrate consistent labeling patterns across different experiment conditions and cell lines.

Using representation-based clustering to derive cysteine ligandability labels for ML. The variability in cysteine ligandability across the ABPP datasets presents a significant challenge for ML. Encouraged by our recent work[47] demonstrating that residue-level representations (also known as per-token embeddings) from ultra-large pLLMs such as ESMC[45,48] encode not only evolutionary information but also structural and local biochemical environments, we leveraged the residue-specific representations to reconcile the labels for liganded and unliganded cysteines. Specifically, we performed clustering analysis on all cysteines in ABPP-quantified proteins using pairwise similarity scores computed from ESMC-derived residue-specific embedding vectors. For each cluster, we identified a representative cysteine as the node exhibiting the highest average similarity to all other cluster members. This approach yielded compact, non-overlapping clusters of cysteine sites with high embedding similarity, each anchored by a representative node used for consensus labeling and subsequent model training. To prevent leakage, clusters were never split across data partitions and all clusters sharing a UID were assigned to the same partition.

To derive high-confidence ligandability labels from the ABPP data, we performed cluster-based label assignment on the full set of representation-based clusters according to varying levels of consensus, defined as a combination of supporting sources and

records (denoted nS–mR). For example, under the 4S–4R criteria, a cluster is labeled pos only if it contains at least four positive records from four different sources. Conversely, a cluster is labeled neg if it contains no positive records and at least four negative records from unique sources. Clusters that did not meet either pos or neg labeling condition were excluded from the training set. To augment coverage, unquantified cysteines in quantified proteins cast default negative votes and inherit a cluster-based label.

Applying the 4S–4R criteria yielded 936 pos and 4,590 neg labeled clusters, covering 143,914 site-level records describing 11,869 distinct cysteine residues across 11,541 unique proteins. Importantly, this approach enabled us to assign labels to 21% of ABPP-unquantified cysteines, classifying 8,273 as pos and 11,665 as neg (Fig. 2c). The cluster representatives were then used for downstream model training. Additional details are provided in the Supplemental Methods, including data validation, representation-level deduplication, and procedures with representation-based criteria to prevent data leakage between training, validation, and external evaluation.

Note, while conventional sequence identity filtering methods, such as CD-HIT,[49] focus on eliminating sequence similarity, they fail to account for functionally convergent or structurally similar ligandable sites that may exist within otherwise divergent proteins. In contrast, clustering based on pLLM representation goes beyond simple sequence comparison by integrating local biochemical, structural, and evolutionary information. Thus, representation-based criteria are more stringent and effective in reducing the risk of data leakage.

Support from 4 sources may be an optimal tradeoff between label confidence and data coverage. To examine how label confidence depends on the number of quantifying sources, we measured variability using Shannon entropy (a measure of labeling uncertainty) of the binary pos/neg distribution for cysteine clusters quantified by n sources. The number of clusters falls sharply as the threshold increases; for example, requiring at least 4 quantifying sources reduces the set to 11,913 clusters compared to 27,410 with a 2-source threshold (Fig. 3a, top). Entropy is high for clusters supported by 2 or 3 sources, reaches its minimum at 4 sources (Fig. 3a, bottom), and then rises again at thresholds ≥ 7, peaking at 12 sources where only $\sim$150 highly heterogeneous clusters remain.

Fig. 3b complements this analysis by showing how positive cysteines accumulate as additional sources are added under different consensus rules (1S–1R, 2S–2R, ..., 5S–5R). At the 4-source threshold, the number of positives normalized by the naïve total (all cysteines reported as liganded in any record) qualitatively plateaus: each additional source contributes relatively few new positives (Fig. 3b, top). Together, these results suggest that requiring evidence from at least 4 supporting sources (defined here as quantifying sources that reported a cysteine as liganded at least once; the 4S–4R rule) provides a practical balance between label confidence and data coverage for annotating ligandable cysteines with the currently available data.

Development of the sequence-only LigCys models. Inspired by our recent demonstration that ESMC-based, sequence-only models can accurately predict residue-specific pK_a values, particularly for cysteines despite limited training data,[47] we developed the LigCys model for predicting cysteine ligandability. The model employs ESMC as the foundational pLLM, using frozen 2,560-dimensional per-residue embeddings passed through a three-layer multilayer perceptron (MLP) classifier trained on the ABPP-derived data described below. Embeddings from layer 76 of ESMC were used (Methods), as preliminary testing showed this layer provided optimal performance for LigCys predictions. Consistent with prior observations,[47,50] different transformer layers capture distinct evolution-

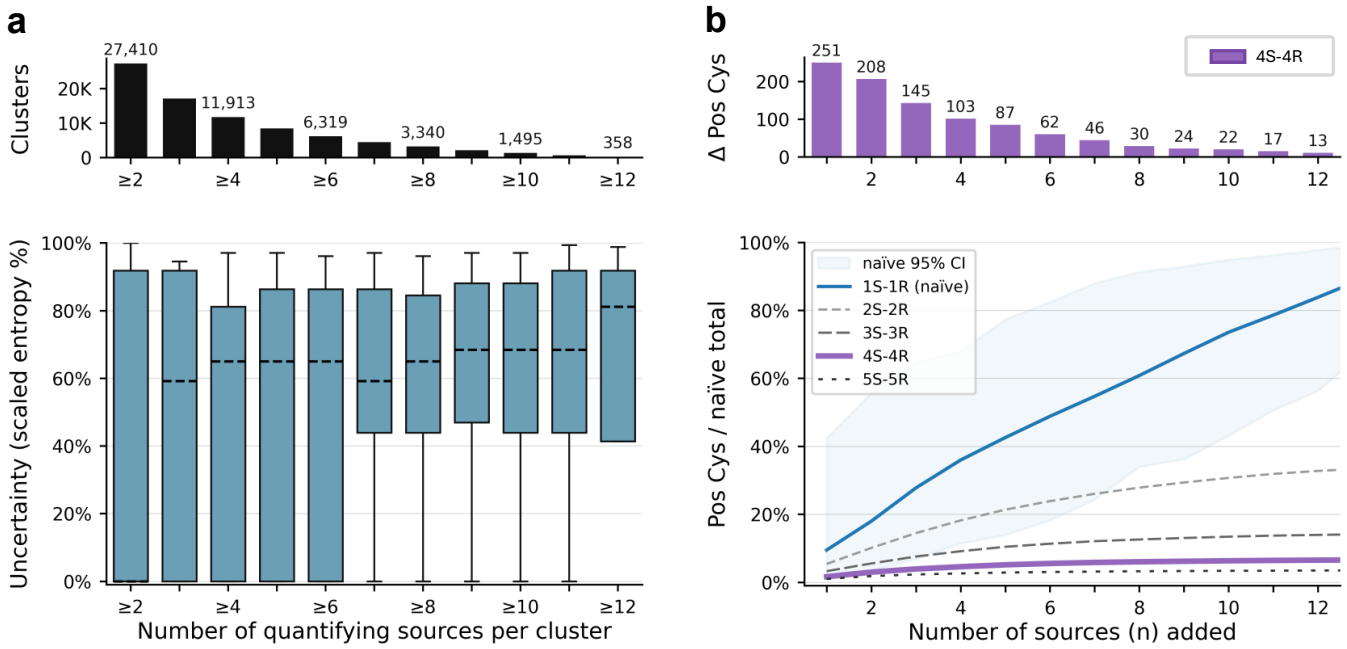

Figure 3: **Analysis of how source threshold affects label variability and data coverage. a.** *Top:* Bar plot showing the number of cysteine clusters supported by at least n independent sources, regardless of the number of underlying records. For clarity, bins corresponding to odd values of n are unlabeled. *Bottom:* Label variability within each cluster, stratified by the number of supporting sources. Variability is quantified using the Shannon entropy of the binary label distribution (p, the fraction of sources labeling a cysteine as positive), scaled to a 0–100% range. Boxes denote the interquartile range; medians are shown as horizontal dashed lines; whiskers span the 10th–90th percentiles. A minimum in entropy is observed at $n \geq 4$, suggesting this threshold maximizes labeling consistency across sources. **b.** *Top:* Incremental gain in the number of positive cysteines when increasing from $n-1$ to n sources. *Bottom:* For each consensus threshold (e.g., 1S–1R through 5S–5R), we randomly sample n sources from a pool of 15, repeat 200 times, and compute the average number of positive cysteines. This is normalized by the naïve total (i.e., all cysteines reported as positive in any single record). The 95% confidence interval is shown for the naïve curve. The 4S–4R consensus curve (highlighted in purple) plateaus after four sources, supporting its use as a data-driven threshold for robust labeling.

ary and structural features at different rates; layer 76 embeddings were therefore also used for LigBind models. We also evaluated pLLM fine-tuning approaches using ESM2,[51] but the resulting models did not outperform the representation-only ESMC approach for cysteine ligandability prediction.

LigCys models trained on the 4S–4R ABPP dataset show the strongest generalization. Due to the highly variable nature of ABPP data, both model training and validation are extremely challenging. To address validation challenges, we used crystallography-validated ligandable cysteines from the LC3D database as an external test set. LC3D cysteines and any representation-based clusters containing these cysteines were removed from the training set to prevent data leakage (Supplemental Methods). Therefore, the LC3D test set is not only orthogonal to the ABPP training set but also provides the most stringent validation for model's ability to generalize.

We evaluated how ABPP data consensus requirement affects the model performance by training LigCys models with ligandability labels derived from increasingly stringent consensus criteria (Table 1). These LigCys models are compared using the standard metrics, including the most pertinent metric Top-1 recovery, which is defined as the probability that the cysteine with the highest predicted classification score in a protein is a true pos given that the protein has at least one pos cys-

Table 1: External evaluation of LigCys models trained using ABPP datasets with varying consensus thresholds[a]

ABPP training data				External LC3D evaluation (%)					
Criteria	**Pos**	**Neg**	**Prot**	**PRC**	**ROC**	**Prec**	**Rec**	**Top-1**	**Top-1***
1S–1R	13430	55441	7051	60.0	72.2	38.4	37.8	40.5	44.5
1S–2R	6078	16791	4216	70.2	80.9	55.9	55.1	56.8	59.6
1S–3R	3396	8600	2829	76.6	85.8	62.3	61.9	65.3	68.2
1S–4R	2094	5048	2004	78.1	86.3	65.3	64.9	68.3	72.9
1S–5R	1418	3354	1495	80.1	87.4	70.0	69.6	71.7	75.9
2S–2R	3467	6838	2617	77.0	84.2	65.3	64.7	66.2	70.3
2S–3R	2198	4765	1883	77.3	85.5	64.3	64.1	66.6	70.3
2S–4R	1450	3326	1404	78.8	86.4	69.6	68.8	70.9	75.1
2S–5R	988	2258	1041	79.5	86.6	71.7	70.9	72.2	76.8
3S–3R	1115	1920	997	78.8	86.6	68.8	68.1	70.0	74.2
3S–4R	867	1583	837	80.2	88.3	68.3	67.5	71.3	75.5
3S–5R	649	1244	669	80.8	87.7	72.6	71.7	73.5	77.6
4S–4R	**384**	**642**	**389**	**81.2**	**87.3**	**74.3**	**73.2**	**75.2**	**79.4**
4S–5R	347	590	358	80.3	86.1	74.3	73.2	74.3	78.9
5S–5R	142	236	150	79.8	85.6	71.3	70.3	72.6	76.8

[a] At a consensus threshold nS–mR, the label assignment requires at least m records from n different sources (see main text). Each model is an ensemble of 200 models (see Supplemental Methods). Performance is measured on the orthogonal LC3D testset on the per-protein basis. Cysteines without covalent ligation are provisionally labeled neg. Precision and recall of predicting ligandable cysteines are calculated with a classification threshold of 0.5. Top-1 recovery refers to the probability that the cysteine with the highest classification score is a true pos given that the protein has at least one pos cysteine. Top-1* refers to the evaluation using LC3D augmented with labels from representation-based clustering (see Supplemental data).

teine. Since LC3D negatives reflect the absence of observed liganding event rather than validated inactivity, AUROC/AUPRC should be interpreted with caution; we therefore emphasize ranked recovery metrics (e.g., Top-1). Confirming the substantial heterogeneity across ABPP data, the 1S–1R model, despite being trained on the largest dataset (13,430 pos and 55,441 neg cysteines), delivers the worst performance, with AUROC, AUPRC, and Top-1 recovery values of 72.2%, 60.0%, and 40.5%, respectively. At the single-source level, performance improves dramatically as the number of required pos records increases, reaching AUROC, AUPRC, and Top-1 recovery of 87.3%, 81.2%, and 71.7%, respectively, despite a ten-fold reduction in training data. A similar trend of significant performance gain with increasing number of pos records is observed at the 2- and 3-source levels. However, at the 4-source level, performance with 5 pos records (4S–5R) is very similar to that with 4 records (4S–4R), which has the AUROC, AUPRC, and Top-1 of 87.3%, 81.2%, and 75.2%, suggesting that model performance may be plateaued. Indeed, the 5-source model (5S–5R) shows decreased performance across all three metrics by 2–5% compared to the 4S–5R model, which may be attributed to the substantial (more than 50%) reduction in training data (only 142 pos and 236 neg cysteines).

This comparison shows that the 4S–4R training set yields the best model performance (Table 1, highlighted row) for generalization to the completely orthogonal external validation data. This observation is consistent with the analyses in Fig. 2 and 3, suggest-

ing that the 4S–4R training set balances label confidence with dataset size for effective generalization. To further assess model performance, we computed Top-1* recovery for the LC3D test set that received additional label assignments through representation-based clustering among themselves (referred to augmented LC3D or LC3D*). The 4S–4R model remains the top-performing one, achieving AUROC, AUPRC, and Top-1* of 87.3%, 81.2%, and 79.4%, respectively.

Enhanced model performance through iterative training data expansion guided by LC3D-based evaluation. Although the 4S–4R dataset achieves superior performance relative to others, its small size may limit the model's generalization capability. To address this limitation, we hypothesized that using LC3D as a validation benchmark would allow us to identify additional cysteines (and proteins) with high-confidence labels from the broader ABPP dataset to expand the training data. To test this hypothesis, we developed an iterative procedure guided by LC3D Top-1 recovery, starting from a truncated version of the high-confidence 4S–4R dataset containing cluster representatives from proteins with both positive and negative cysteines. In this truncated version (denoted 4S–4R^t), 23 proteins quantified by 10 or more sources with at least one pos and one neg cysteine are withheld for model evaluations. The candidate training pool included cysteines from all 4S–4R^t clusters, extending beyond the baseline to encompass non-representative members as well as proteins supported exclusively by positive or exclusively by negative evidence.

In iterations 1–3, we generated 100 batches, each consisting of 80, 100, and 150 randomly selected cysteines from the candidate pool, respectively. From iteration 4 onward, we generated 100 batches of 175 randomly selected cysteines. In each iteration, batches were evaluated by incorporating them into the training set and training ensembles of 24 models (6 data splits with 4 random seeds per split). The best-performing batch was selected for further validation using 10 new random splits with 20 models per split, with Top-1 recovery serving as the evaluation metric. If Top-1 recovery exceeded the previous iteration's baseline, the best batch was accepted.

Through six iterations, Top-1 recovery increased from 70.9% to 79.9% on LC3D and from 75.1% to 84.5% on LC3D*, accompanied by steady improvements in AUROC from 85.5% to 88.9% and AUPRC from 78.1% to 84.2% (Table 2). In iteration 7, no further gain was observed, indicating that performance had plateaued under this random-batch expansion scheme. Hereafter we will call the Iter6 model LigCys-S. Relative to the baseline 4S–4R^t dataset, the expanded training set nearly tripled in size to 1,099 proteins, driven by over twofold increase in negative cysteines (1,345) and a nearly 30% increase in positives (441). Using the same data, we trained a structure-aware model (6-SA) to test if adding additional structural features would enhance predictive power (Table 2, row 6-SA). Interestingly, along with other calculated metrics, the Top-1 recovery of LC3D liganded cysteines is decreased by 1.7 %. This suggests that the evolutionary driven pLLM-based representations capture the majority of information pertinent to cysteine ligandability.

Enhanced model performance through iterative training data expansion guided by ABPP-based evaluation. As random sampling in the LC3D-guided procedure may leave parts of the candidate pool unexplored, we hypothesized that a systematic, cross-validation (CV) driven approach could more efficiently identify additional high-confidence data. To test this hypothesis, we developed an expansion strategy guided by CV performance on the ABPP data, starting from the same baseline dataset (4S–4R^t) as in the LC3D-guided expansion. The same candidate data pool was also used.

At each iteration, candidate batches of fixed composition (25 positives and the remainder negatives) were assembled, with batch

Table 2: Performance metrics for iterative training data expansion guided by LC3D-based evaluation (top block) and ABPP-based cross validation (bottom block).

ABPP training data[a]				ABPP valid. (%)[b]		External LC3D evaluation[c] (%)					
LC3D-guided expansion											
Iter	**Pos**	**Neg**	**Prot**	**PRCE**	**ROC**	**PRC**	**ROC**	**Prec**	**Rec**	**Top-1**	**Top-1***
0	354	574	366	74	74.6	78.1	85.5	70.5	69.6	70.9	75.1
0-SA	354	574	366	79	76.5	80.1	86.9	71.3	70.5	72.6	76.8
1	359	649	442	92	78.2	79.5	85.8	71.3	70.3	72.2	76.8
2	369	739	534	94	76.1	80.4	86.8	73.2	73.4	73.5	78.1
3	384	874	674	92	74.2	81.8	87.8	73.9	72.8	74.8	79.8
4	401	1032	827	115	77.5	82.9	88.0	78.2	76.9	78.2	82.8
5	425	1186	966	139	79.4	83.0	87.8	76.9	75.6	78.2	82.8
6	**441**	**1345**	**1099**	**141**	**78.2**	**84.2**	**88.9**	**77.3**	**77.7**	**79.9**	**84.5**
6-SA	441	1345	1099	128	76.7	83.8	88.7	76.9	75.6	78.2	83.2
ABPP-guided expansion											
Tour[d]	**Pos**	**Neg**	**Prot**	**PRCE**	**ROC**	**PRC**	**ROC**	**Prec**	**Rec**	**Top-1**	**Top-1***
0	354	574	366	74	74.6	78.1	85.5	70.5	69.6	70.9	75.1
1	378	643	452	93	80.2	80.1	86.8	72.6	71.7	73.1	77.3
2	468	967	817	122	82.1	81.5	88.3	78.1	72.1	74.4	79.4
3	555	1432	1258	148	82.6	80.1	87.5	71.3	70.3	71.8	76.0
4	**645**	**1998**	**1744**	**187**	**83.5**	**79.3**	**87.1**	**69.8**	**69.2**	**70.5**	**76.0**
4-SA	645	1998	1744	178	80.9	80.5	87.7	69.8	69.2	71.7	76.8
Blended[e]	-	-	-	-	-	83.2	88.4	75.4	74.5	76.9	85.5

[a] In the LC3D-guided data expansion, Top-1 recovery of LC3D liganded cysteines was used to select candidate data batches (complete data in Supplemental Data). The total number of pos, neg cysteines, and distinct proteins in the training set is given after each iteration. The start of the expansion (Iter 0) used the 4S–4R^t dataset, which excludes 23 proteins quantified by 10 sources from the original 4S–4R dataset. [b] Model validation metrics are AUROC (ROC) and AUPRC enrichment (PRCE), defined as $(\text{AUPRC} - \text{random})/\text{random}$. Here random represents the AUPRC of a random classifier (% positive samples in the dataset). [c] For external evaluation against the LC3D dataset, precision (Prec) and recall (Rec) were calculated using the default classification threshold of 0.5. Top-1 and Top-1* were calculated using LC3D and LC3D* (label augmented by clustering) datasets, respectively. [d] In the ABPP-guided data expansion, AUPRC enrichment (PRCE) was used to select candidate data batches. Tour denotes the number of successive tournaments (effectively iterations 1, 5, 9, and 13; see Supplemental Data). All other details follow the notes above. [e] Blended denotes ensemble prediction using both Iter 6 and Tour 4 models.

size set to 100 cysteines in early iterations and increased to 175 as the pool expanded. Batches were prioritized by ensemble uncertainty, and a batch was admitted only if it significantly exceeded the AUPRC enrichment (PRCE) of the previous iteration baseline in the CV. PRCE (enrichment of AUPRC over random guess) was used due to the variable pos:neg Cys ratio in each iteration. To broaden the search, advancing batches were further refined by simple genetic algorithm steps, preserving the best performers while generating crossover and mutation variants.

Expansion proceeded through four tournaments, each comprising four iterations. Here, a "tournament" refers to a series of evaluations in which candidate batches compete against the baseline: top performers

are retained and recombined, and only the strongest batch is accepted at each iteration of the tournament. Expansion terminated when the final tournament ended without any accepted batch. This acceptance scheme yielded steady performance gains: Top-1 recovery stayed constant at 70.9% to 70.5% on LC3D and at 75.1% to 76.0% on LC3D*, with corresponding increases in AUROC and AUPRC (Table 2, under ABPP-guided expansion). Hereafter we will refer to the Tour4 model as LigCys-A. The final expanded set comprised 1,744 proteins and 2,633 cysteine residues, representing a 1.8-fold increase in positives and a 3.4-fold increase in negatives relative to the 4S–4R^t baseline. Performance gains plateaued after the fourth tournament, indicating that the available candidate pool had been exhausted. As in the LC3D-guided expansion, we also trained a structure-aware model using the final expanded data. The resulting model (4-SA) has a small increase of 1.2% for AUPRC and Top-1 recovery compared to the sequence-only model.

Beyond the LC3D test, we evaluated the LigCys-S and LigCys-A models against the ABPP hold-out set comprised of 23 proteins quantified by 10 or more sources, each containing at least one positive and one negative cysteine. From this hold-out set, we created 9 test sets using the nS-nR consensus thresholds (n=1,...,9). As expected, the number of labeled cysteines decreases from 1S-1R ($\sim$300 cysteines) to just under 100 at 4S-4R and 7 at 9S-9R (Supplemental Fig. S2). The AUPRC of both models increases with the consensus level of the test set, from $\sim$68% (67 for LigCys-A model) at 1S-1R to 73% (75 for LigCys-A model) at 4S-4R and nearly 100% at 9S-9R (Supplemental Fig. S2). Similarly, the model recall steadily increases from below 40% at 1S-1R to 100% at 9S-9R (Supplemental Fig. S2). These improvements stem from increased TPs and decreased FNs with growing consensus level (Supplemental Fig. S2). The analysis demonstrates that the models perform better on consistently liganded cysteines, corroborating our initial findings (Table 1). Consistent with the LC3D test, the ABPP hold-out test suggests that LigCys-S and LigCys-A models perform similarly, with LigCys-S showing slightly fewer false positives when predicting highly ligandable cysteines.

Development of an ML-ready LigBind3D database and the sequence-only LigBind models for predicting reversible ligand binding residues. Since truly ligandable cysteines are located at or near a reversible binding pocket, we developed sequence-only LigBind models (LigBind-seq) to predict reversible ligand-binding residues complementary to cysteine ligandability predictions. To develop the LigBind models, we created an ML-ready database called LigBind3D based on BioLiP2,[52] an updated manually curated dataset of experimental structures from biologically relevant protein–ligand complexes in the PDB. BioLiP2[52] is significantly larger than commonly used datasets for protein-ligand affinity predictions (e.g., BindingDB[53]) which contain only protein–ligand interactions with experimentally determined binding affinities.

We selected BioLiP2 entries of monomeric proteins containing small molecules with molecular weights between 150–600 Da and applied a custom structural annotation pipeline (`PickPocket`, unpublished) to validate biologically relevant protein-ligand interactions. Following the various filtering steps, we reconstructed the full protein sequence directly from the atomic coordinate file. Ligand-binding residues were labeled based on a 4.5 Å heavy-atom distance cutoff. These residues were converted into structured, site-level records. Details are given in Supplemental Methods. The final LigBind3D database comprises 687,712 site-level records across 1,998 unique proteins. This database offers a direct structural perspective on reversible protein-ligand interactions, complementing the LigCysABPP and LC3D databases. In contrast to LigCys3D, which restricts to cysteines, LigBind3D labels every residue in a protein, reflecting the broader diversity of reversible protein-ligand interactions. Sim-

ilar to the treatment of LigCysABPP data, representation-based clustering was applied to the LigBind3D data for creating the training, validation, and test datasets.

The LigBind-seq model employs a simpler single-layer feedforward network architecture that processes fixed 2,560-dimensional per-token representations extracted from ESMC layer 76. Preliminary screening showed that the simpler architecture was sufficient to capture ligand-binding patterns from sequence alone. In addition to the sequence-only models, we also developed structure-aware LigBind models (LigBind-SA) by incorporating 3D geometric features through a pretrained transformer model derived from the PeSTo architecture[54] alongside the pLLM embeddings. Details are given in Supplemental Methods.

In the hold-out test, LigBind-seq achieves AUROC, AUPRC, and Top-10 scores of 93.9%, 74.5%, and 78.1%, respectively, while LigBind-SA attains 95.1%, 77.5%, and 80.1% (Table 3). The modest performance improvements of LigBind-SA relative to LigBind-seq indicate that the ESMC-derived representations provide the majority of the model's predictive power.

Table 3: Performance of the sequence-only and structure-aware LigBind models in predicting reversible ligand-binding residues[a]

Metrics (%)	LigBind-seq	LigBind-SA
AUROC*	93.9	95.1
AUPRC*	74.5	77.5
Prec*	72.1	71.1
Rec*	66.4	71.5
Top-5*	79.2	81.2
Top-10*	78.1	80.1
Top-20*	76.6	79.5

[a] All models were trained and evaluated* using cluster-propagated labels derived via representation-based clustering. Performance metrics reflect a post hoc ensemble of 200 models (Supplemental Methods). Per-residue ligandability probabilities were averaged across all 200 models and thresholded at 0.5. All metrics were computed on a per-protein basis using the held-out test set of 129 proteins.

LigBind predictions offer complementary validation for LigCys predictions. Covalent ligandability of cysteines requires proximity to a reversible binding pocket. Therefore, accurate LigBind predictions not only provide reversible pocket information but may also serve as complementary validation for LigCys predictions. To test this complementary relationship, we applied both LigBind and LigCys models to the LC3D test set with proteins in the LigBind training set excluded. We then calculated the radial distribution function (RDF) of predicted ligand-binding residues surrounding TP, TN, FP, and FN cysteines (Fig. 4). Both LigBind-seq and LigBind-SA models were used. Except for TNs, all RDFs exhibit a peak around 4–6 Å. Notably, both LigBind-seq and LigBind-SA RDFs for TPs show the highest peak, consistent with the expectation that ligandable cysteines are proximal to reversible binding pockets. The peak of the LigBind-SA RDF (dashed) is a bit higher compared to the LigBind-seq RDF (solid), suggesting that incorporating structural features enhances pocket prominence. Reassuringly, both LigBind-seq and LigBind-SA RDFs for TNs lack peaks, consistent with the expectation that unligandable cysteines are not located near ligand-binding pockets. Remarkably, the LigBind-seq RDF for FNs exhibits the second highest peak, suggesting that LigBind-seq predictions could help identify overlooked ligandable cysteines based on pocket proximity. Conversely, the LigBind-seq and LigBind-SA RDFs for FPs display very low peaks, suggesting that incorrectly predicted ligandable cysteines can be filtered out based on their lack of proximal binding pockets. These results demonstrate that integrating LigCys and LigBind predictions within a unified framework would further enhance the reliability of covalent ligandability assessment.

Development of the AiPP platform for sequence-based predictions of reversible and covalent ligand binding sites across the proteome. Building on the LigCys and

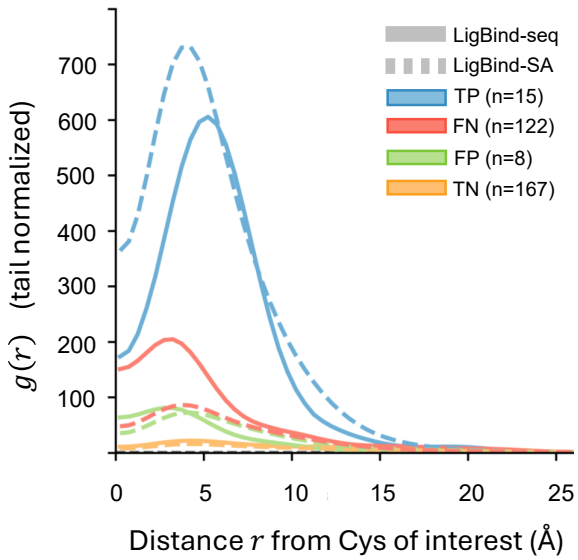

Figure 4: **Radial distribution function of LigBind-predicted ligand-binding residues around cysteine sites, stratified by the LigCys prediction outcome (TP, TN, FP, FN).** Distance is the minimum all-atom Euclidean distance between the cysteine of interest and each predicted ligand-binding residue. Radial distribution function $g(r)$ is normalized by the mean shell density over the last 20% of the distance range. Solid and dashed lines represent the LigBind-seq and LigBind-SA predictions, respectively, on the pruned LC3D test set that excludes proteins in the LigBind training set.

LigBind models, we developed the Artificial Intelligence Protein Profiling (AiPP) platform as a unified, sequence-based framework for proteome-wide ligandable site identification (Fig. 5). AiPP integrates reversible and covalent ligand-binding predictions as well as a disorder predictor to support rational drug discovery. The LigBind module detects residues that spatially stabilize ligand binding; the LigCys module identifies cysteines susceptible to modification by a covalent ligand; and the RIDA module (based on the RIDAO server[56]) predicts per-residue disorder and molecular recognition features (MoRFs) that can fold upon binding. Together, these predictions map ligandable binding sites capable of retaining cysteine-reactive compounds sufficiently long for covalent bond formation—guiding targeted covalent drug design, while simultaneously identifying dynamic or induced-fit binding regions amenable to PPI inhibitor development. Additionally, the recently developed KaML-ESM[47] and KaML-CBtree[57] models are incorporated to offer reactivity information for cysteines of interest based on sequence only or structure.

Starting from a protein sequence, AiPP generates a predicted 3D structure (via a local version of ESM3[48]), a structural and functional disorder annotation (via a local version of RIDAO server[56]), and the residue-specific representations (from the pLLM ESMC[45]). The ESMC-based representations are processed by LigBind and LigCys modules to generate predictions of reversible and cysteine-directed covalent ligand binding sites, respectively. LigCys generates predictions from the LigCys-S and LigCys-A models, as well as a blended prediction that ensembles results from both models. Optionally, structure-aware versions of the predictions are provided. Results are visualized on the ESM3 predicted structure with potential MoRFs. Optionally, the predicted structure is processed by the pretrained transformers derived from the PeSTo architecture[54] to generate geometric representations, which are then integrated with the ESMC-based representations as well as hand-engineered features to enhance predictions by LigBind and LigCys. Additionally, cysteine reactivity (using pK_a as a proxy) and solvent accessibility (buried fraction) are annotated to inform ligand design.

Evaluation on the consistently and heterogeneously liganded cysteines from a recent ABPP study.[31] A recent ABPP study using over 400 cancer cell lines demonstrated that while a majority of the quantified cysteines were consistently liganded across cell lines, a subset of cysteines showed varied ligandability depending on the cell line (referred to as heterogeneous cysteine ligandability).[31] We made predictions for 14 proteins containing the consistently and heterogeneously liganded cysteines featured in the publication[31] (DrugMap dataset) but absent in the LigCys training set. Note, although LigCysABPP database does not contain the DrugMap dataset, there is a significant overlap in proteins/cysteines. LigCys predicted

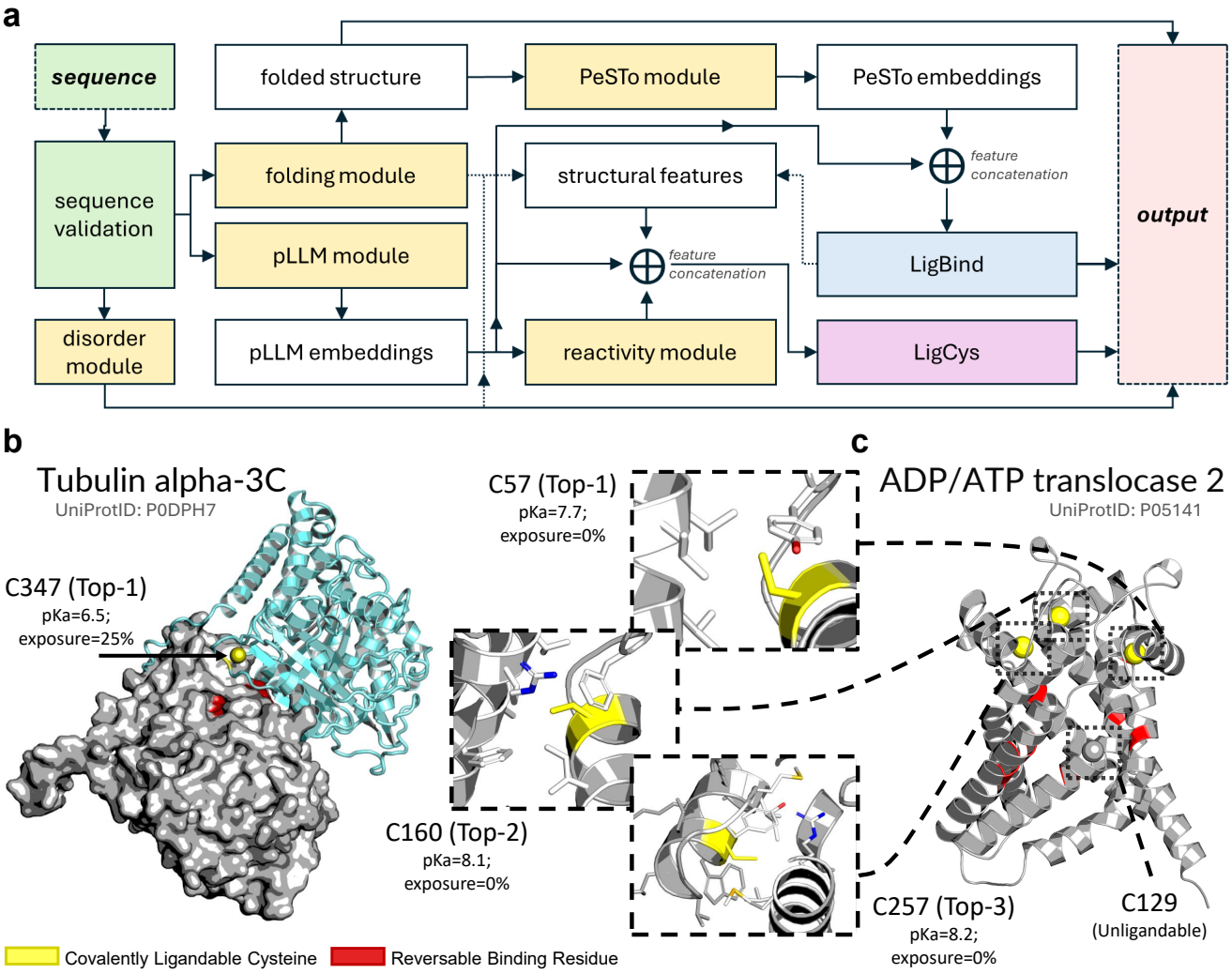

Figure 5: **Architecture of the AI protein profiler (AiPP) for annotating covalent and reversible ligand binding sites given a protein sequence. a.** Starting from a protein sequence, AiPP generates a predicted 3D structure (ESM3), structural and functional disorder annotations (RIDA), and ESMC-based representations. These representations are processed by three specialized modules: LigCys, LigBind, and KaML-ESM to predict covalent, reversible ligand binding residues, cysteine reactivities, respectively. Results are mapped onto the ESM3 predicted structure. Optionally, the predicted structure is processed by a transformer model (PeSTo) to generate geometric representations, which are then integrated with the pLLM-based embeddings as well as hand-engineered features to generate LigCys-SA and LigBind-SA models. **b.** For TUBA3C, LigCys predicted C347 as Top-1 ligandable, supporting the finding that it is consistently liganded across cell lines.[31] C347 is at the binding interface with β-tubulin in the adjacent dimer of the protofilaments that comprise microtubule structures. For visualization, we added a surface rendering of β-tubulin based on the electron diffraction model (PDB ID: 1TUB).[55] **c.** For SLC25A5, LigCys predicted C57, C160, C257, and C129 as Top-1, Top-2, and Top-3 ligandable, respectively, while C129 was predicted as unligandable, consistent with the order of engagement frequency across cell lines.[31]

tubulin alpha-3C (TUBA3C, UID P0DPH7) Cys347 and SOX10 (UID P56693) Cys71 as Top-1 ligandable (Fig. 5b), corroborating the consistent engagement across different cancer cell lines.[31] Although this comparison is limited, the data suggest that LigCys models can confidently predict consistently ligandable cysteines.

Based on the electron microscopy model of α/β-tubulin dimer,[55] Cys347 is positioned at the binding interface with β-tubulin from the adjacent dimer in the protofilaments that comprise the microtubule structures. Given the disease relevance of tubulin proteins, the convergence of computational predictions and experimental evidence for Cys347 ligandability suggests potential drug design opportunity for disrupting or stabilizing this protein-

protein interaction (PPI). SOX10 transcription factor is a major dependency of melanoma cells.[31] Notably, the compelling ligandability of Cys71, which is located at the dimerization domain, facilitated a proof-of-concept ligand design that stabilizes the PPI, leading to the disruption of melanoma transcriptional signaling.[31]

Among the heterogeneously liganded cysteines[31] not in the LigCys training set, LigCys predictions vary. While PCBP1 (Q15365) Cys109, LARS1 (Q9P2J5) Cys70, and RPS3A (P61247) Cys201, are predicted as non-ligandable, SLC25A5 (P05141) Cys257 and ACAT1 (A0A5F9ZI66) Cys413 were each predicted as the Top-2 ligandable cysteine in their respective proteins. Interestingly, closer examination of our prediction and the published ABPP data[31] for SLC25A5 revealed that the LigCys ranking of the 3 predicted ligandable cysteines (Fig. 5c) is identical to the order of the positive engagement frequency across cell lines. Note, a positive engagement is defined by $\geq$60% engagement, corresponding to the commonly used CR threshold of 4.[31] While this level of agreement may be anecdotal given our limited comparison, the data suggest that LigCys models may already capture some information about heterogeneous ligandability through the prediction ranking. While our LC3D-based evaluations focused solely on the Top-1 predictions, users can explore additional candidates by examining cysteines ranked 2nd, 3rd, or lower in their ligandability.

Concluding Discussion

We present the first unified AI platform to enable sequence-based, scalable annotation of reversible and covalent ligand-binding sites across the proteome. This platform is enabled through the new ML-ready databases and methodological advances that address several critical challenges in proteome-wide ML.

By incorporating positive, negative, and unquantified ABPP records from 15 publications, the LigCysABPP database substantially exceeds the scope and comprehensiveness of existing ABPP databases, including CysDB,[43] DrugMap,[31] and TopCysteineDB.[32] LigCysABPP comprises 608,898 quantified and 94,237 unquantified cysteine records of 10,649 unique proteins. Of these proteins, 6,502 contain at least 1 liganded cysteine (a total of 15,559 liganded cysteines) while 4,147 contain only unliganded and unquantified cysteines. Building upon BioLip2[52] the LigBind3D database offers more comprehensive coverage of protein-ligand interactions than commonly used binding databases such as BindingDB.[53] LigBind3D contains annotations for reversible ligand-binding residues across 1,998 unique proteins derived from co-crystal structures. Together with LC3D (a refined database of covalently liganded cysteines evidenced by co-crystal structures), LigCysABPP and LigBind3D provide solid foundations for developing ML models aimed at proteome-wide discovery of covalent (LigCys) and reversible (LigBind) ligand-binding sites.

A major bottleneck in developing ML models for predicting ligand binding residues is experimental label variability (e.g., conflicting annotations for ligandable cysteines) and incompleteness (lack of high-confidence negatives for covalent or reversible ligand-binding residues). We addressed these issues by developing a pLLM representation-based clustering approach to reconcile and augment labels. To develop robust LigCys models from ABPP data, we established a consensus analysis framework evaluating models trained on data at varying consensus thresholds, using LC3D as an orthogonal validation dataset to identify the most reliable baseline. Building from this optimized baseline, we implemented two iterative approaches – LC3D- (orthogonal validation) and ABPP-guided (inherent validation) – to systematically increase the training set size and diversity, achieving substantial improvements in model performance and generalization.

The structural requirement of ML models significantly limits practical applications.

We addressed this constraint by leveraging state-of-the-art pLLM ESMC[45] to develop LigCys and LigBind models capable of rapid, sequence-based prediction of covalent and reversible ligand-binding sites for proteome-scale exploration. The LigCys models achieved 79.9% Top-1 recovery of all known liganded cysteines from co-crystal structures. The LigBind models demonstrated strong performance on the hold-out test set of 129 proteins, achieving an AUROC of 93.9%, AUPRC of 74.5%, and Top-10 recovery rate of 78.1%.

LigCys and LigBind models are integrated into AiPP, a unified AI platform designed for proteome-scale annotation of covalent and reversible ligand-binding sites. Additionally, AiPP incorporates several complementary predictive tools: a meta disorder predictor RIDA (based on RIDAO[56]) for identifying MoRFs critical to PPI drug design, a cysteine reactivity predictor[47,57] for additional functional studies, and ESM3[48] for protein structure prediction. The integration of structural information (LigBind-SA) enhances binding site predictions and provides users with visualization capabilities. Importantly, AiPP can be systematically improved (e.g., by incorporating continuously growing data from ABPP experiments) and expanded (e.g., by adding prediction capabilities for lysine and other covalently ligandable residues).

Current work has several caveats that can be addressed in the future. Although limited comparison to the recent DrugMap work[31] suggest that LigCys models may already capture some information about heterogeneous ligandability through the prediction ranking system, current LigCys models are currently agnostic to cellular context, e.g., redox conditions, post-translational modifications, and mutational states. Two recent ABPP studies revealed that phosphorylation[33] and specific cancer-cell environment[31] can significantly modulate cysteine reactivity and/or ligandability in a subset of cysteines. The underlying mechanisms may involve phosphorylation-induced protein conformational changes, shifts in cellular redox state, and accumulated cysteine-proximal mutations in malignant cells.[31,33] Molecular dynamics (MD) simulations[58,59] have demonstrated that the nucleophilicity of cysteines (and other residues such as lysines) can be conformation-dependent, with kinases providing clear examples of how reactivity varies between functionally relevant conformational states (e.g., DFG-in vs. DFG-out forms). We envision that the contextual information can be added to the training data and at inference stage to provide proteoform-specific predictions in the future. This enhancement may resolve the divergent performance observed between models trained on datasets expanded using LC3D- vs. ABPP-guided strategies. Another limitation of current AiPP is the LigBind model which was trained using a database consisting of monomer proteins. In addition to addressing these limitations, we also envision coupling AiPP with ligand docking and generation models to advance proteome-wide drug design. Finally, the representation-based label analysis and augmentation approaches developed in this work are broadly applicable for ML-experiment integration, establishing a generalizable framework for developing proteomics-derived ML models across diverse applications.

Supplemental Materials

Supplemental Methods contains detailed procedures for constructing the LigCysABPP, LC3D, and LigBind3D databases; pLLM-based data clustering, label assignment and reconciliation; LigCys and LigBind model architectures, training, and validation; and LigCys training data expansion. Supplemental Figures contains a more detailed analysis accompanying Figure 1; model performance analysis using the ABPP hold-out set.

Supplemental Data

A Supplemental Data Excel file contains the following 20 sheets:

Benchmarking studies: ESMC representation (SD2), consensus (SD3), source ablation (SD4), source prioritization (SD5), option-ablation (SD20) and split-level noise (SD21).
Data extraction logs: ABPP extraction log (SD6), log of ambiguous and multivalue entry splitting (SD7), ABPP statistical report (SD8), ABPP orphaned records (dropped due to lack of unambiguous support) (SD9).
Sequence and clustering records: validated UID sequences in FASTA format (SD10), UID deduplication log (SD11), precluster records (SD12), cysteine (Cys) clusters report (SD13).
Production distillation: 4S–4R distillation report (SD14).
Support reports: UID-level (SD15), ROI-level (SD16), and representative-level (SD17).
ABPP hold-out set details: UID-level (SD18) and ROI-level (SD19).
Note: Production runs refer to combined ABPP and LC3D data processed in a single distillation workflow, as illustrated in Fig. 1a.

Data Availability

Databases and models are available at `https://github.com/JanaShenLab/AiPP/`, which also contains a link to the corresponding web server `https://aipp.computchem.org/`. Two data files accompany this manuscript: the Supplemental Data spreadsheet (SD2–SD21) and the LigCys-ABPP_v1.xlsx spreadsheet; both are mirrored with code, models, and data in the repository.

Acknowledgments

Research reported in this publication was supported by the National Institute Of General Medical Sciences (R35GM148261) and National Cancer Institute (R01CA256557) of the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We thank Benjamin Cravatt from the Scripps Research Institute for insightful comments and discussion. We thank Chetan Mishra and Marius Wiggert from Evolutionary Scale for facilitating the usage of ESM Cambrian and ESM3. This work was supported by an EvolutionaryScale compute grant.

References

(1) Aebersold, R. et al. How Many Human Proteoforms Are There? *Nat. Chem. Biol.* **2018**, *14*, 206–214.

(2) Wishart, D. S. et al. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074–D1082.

(3) Liu, Y.; Patricelli, M. P.; Cravatt, B. F. Activity-Based Protein Profiling: The Serine Hydrolases. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 14694–14699.

(4) Greenbaum, D.; Medzihradszky, K. F.; Burlingame, A.; Bogyo, M. Epoxide Electrophiles as Activity-Dependent Cysteine Protease Pro¢ling and Discovery Tools. *Chem. Biol.* **2000**, *7*, 569–581.

(5) Chan, W. C.; Sharifzadeh, S.; Buhrlage, S. J.; Marto, J. A. Chemoproteomic Methods for Covalent Drug Discovery. *Chem. Soc. Rev.* **2021**, *50*, 8361–8381.

(6) Cravatt, B. F. Activity-Based Protein Profiling – Finding General Solutions to Specific Problems. *Isr. J. Chem.* **2023**,

(7) 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*, 790–795.

(8) 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.

(9) Roberts, A. M.; Miyamoto, D. K.; Huffman, T. R.; Bateman, L. A.; Ives, A. N.; Akopian, D.; Heslin, M. J.; Contreras, C. M.; Rape, M.; Skibola, C. F.; Nomura, D. K. Chemoproteomic Screening of Covalent Ligands Reveals UBA5 As a Novel Pancreatic Cancer Target. *ACS Chem. Biol.* **2017**, *12*, 899–904.

(10) Weerapana, E.; Speers, A. E.; Cravatt, B. F. Tandem Orthogonal Proteolysis-Activity-Based Protein Profiling (TOP-ABPP)—a General Method for Mapping Sites of Probe Modification in Proteomes. *Nat. Protoc.* **2007**, *2*, 1414–1425.

(11) Wang, C.; Weerapana, E.; Blewett, M. M.; Cravatt, B. F. A Chemoproteomic Platform to Quantitatively Map Targets of Lipid-Derived Electrophiles. *Nat. Methods* **2014**, *11*, 79–85.

(12) Bar-Peled, L. et al. Chemical Proteomics Identifies Druggable Vulnerabilities in a Genetically Defined Cancer. *Cell* **2017**, *171*, 696–709.e23.

(13) Vinogradova, E. V. et al. An Activity-Guided Map of Electrophile-Cysteine Interactions in Primary Human T Cells. *Cell* **2020**, *182*, 1009–1026.e29.

(14) Cao, J.; Boatner, L. M.; Desai, H. S.; Burton, N. R.; Armenta, E.; Chan, N. J.; Castellón, J. O.; Backus, K. M. Multiplexed CuAAC Suzuki–Miyaura Labeling for Tandem Activity-Based Chemoproteomic Profiling. *Anal. Chem.* **2021**, *93*, 2610–2618.

(15) Kuljanin, M.; Mitchell, D. C.; Schweppe, D. K.; Gikandi, A. S.; Nusinow, D. P.; Bulloch, N. J.; Vinogradova, E. V.; Wilson, D. L.; Kool, E. T.; Mancias, J. D.; Cravatt, B. F.; Gygi, S. P. Reimagining High-Throughput Profiling of Reactive Cysteines for Cell-Based Screening of Large Electrophile Libraries. *Nat. Biotechnol.* **2021**, *39*, 630–641.

(16) Yan, T.; Desai, H. S.; Boatner, L. M.; Yen, S. L.; Cao, J.; Palafox, M. F.; Jami-Alahmadi, Y.; Backus, K. M. SP3-FAIMS Chemoproteomics for High-Coverage Profiling of the Human Cysteinome**. *ChemBioChem* **2021**, *22*, 1841–1851.

(17) 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.

(18) Yang, F.; Jia, G.; Guo, J.; Liu, Y.; Wang, C. Quantitative Chemoproteomic Profiling with Data-Independent Acquisition-Based Mass Spectrometry. *J. Am. Chem. Soc.* **2022**, *144*, 901–911.

(19) Yan, T.; Boatner, L. M.; Cui, L.; Tontonoz, P. J.; Backus, K. M. Defining the Cell Surface Cysteinome Using Two-Step Enrichment Proteomics. *JACS Au* **2023**, *3*, 3506–3523.

(20) Burton, N. R.; Polasky, D. A.; Shikwana, F.; Ofori, S.; Yan, T.; Geiszler, D. J.; Veiga Leprevost, F. D.; Nesvizhskii, A. I.; Backus, K. M. Solid-Phase Compatible Silane-Based Cleavable Linker Enables Custom Isobaric Quantitative Chemoproteomics. *J. Am. Chem. Soc.* **2023**, *145*, 21303–21318.

(21) Koo, T.-Y.; Lai, H.; Nomura, D. K.; Chung, C. Y.-S. N-Acryloylindole-alkyne (NAIA) Enables Imaging and Profiling New Ligandable Cysteines and Oxidized Thiols by Chemoproteomics. *Nat. Commun.* **2023**, *14*, 3564.

(22) Burton, N. R.; Backus, K. M. Functionalizing Tandem Mass Tags for Streamlining Click-Based Quantitative Chemoproteomics. *Commun. Chem.* **2024**, *7*, 80.

(23) Njomen, E.; Hayward, R. E.; DeMeester, K. E.; Ogasawara, D.; Dix, M. M.; Nguyen, T.; Ashby, P.; Simon, G. M.; Schreiber, S. L.; Melillo, B.; Cravatt, B. F. Multi-Tiered Chemical Proteomic Maps of Tryptoline Acrylamide–Protein Interactions in Cancer Cells. *Nat. Chem.* **2024**, *16*, 1592–1604.

(24) Tian, C.; Sun, L.; Liu, K.; Fu, L.; Zhang, Y.; Chen, W.; He, F.; Yang, J. Proteome-Wide Ligandability Maps of Drugs with Diverse Cysteine-Reactive Chemotypes. *Nat. Commun.* **2025**, *16*, 4863.

(25) Biggs, G. S. et al. Robust Proteome Profiling of Cysteine-Reactive Fragments Using Label-Free Chemoproteomics. *Nat. Commun.* **2025**, *16*, 73.

(26) Abegg, D.; Frei, R.; Cerato, L.; Prasad Hari, D.; Wang, C.; Waser, J.; Adibekian, A. Proteome-Wide Profiling of Targets of Cysteine Reactive Small Molecules by Using Ethynyl Benziodoxolone Reagents. *Angew. Chem. Int. Ed.* **2015**, *54*, 10852–10857.

(27) Shannon, D. A.; Weerapana, E. Covalent Protein Modification: The Current Landscape of Residue-Specific Electrophiles. *Curr. Opin. Chem. Biol.* **2015**, *24*, 18–26.

(28) Long, M. J. C.; Aye, Y. Privileged Electrophile Sensors: A Resource for Covalent Drug Development. *Cell Chem. Biol.* **2017**, *24*, 787–800.

(29) Wang, S.; Tian, Y.; Wang, M.; Wang, M.; Sun, G.-b.; Sun, X.-b. Advanced Activity-Based Protein Profiling Application Strategies for Drug Development. *Front. Pharmacol.* **2018**, *9*.

(30) White, M. E.; Gil, J.; Tate, E. W. Proteome-Wide Structural Analysis Identifies Warhead- and Coverage-Specific Biases in Cysteine-Focused Chemoproteomics. *Cell Chem. Biol.* **2023**, *30*, 828–838.e4.

(31) Takahashi, M. et al. DrugMap: A Quantitative Pan-Cancer Analysis of Cysteine Ligandability. *Cell* **2024**, *187*, 2536–2556.e30.

(32) Bonus, M.; Greb, J.; Majmudar, J. D.; Boehm, M.; Korczynska, M.; Nazemi, A.; Mathiowetz, A. M.; Gohlke, H. *TopCysteineDB*: A Cysteinome-wide Database Integrating Structural and Chemoproteomics Data for Cysteine Ligandability Prediction. *J. Mol. Biol.* **2025**, *437*, 169196.

(33) Kemper, E. K.; Zhang, Y.; Dix, M. M.; Cravatt, B. F. Global Profiling of Phosphorylation-Dependent Changes in Cysteine Reactivity. *Nat. Methods* **2022**, *19*, 341–352.

(34) Mann, M.; Kulak, N. A.; Nagaraj, N.; Cox, J. The Coming Age of Complete, Accurate, and Ubiquitous Proteomes. *Mol. Cell* **2013**, *49*, 583–590.

(35) Gao, M.; Moumbock, A. F. A.; Qaseem, A.; Xu, Q.; Günther, S. CovPDB: A High-Resolution Coverage of the Covalent Protein–Ligand Interactome. *Nucleic Acids Res.* **2022**, *50*, D445–D450.

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

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

(38) Jumper, J. et al. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.

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

(40) Reimer, B. M.; Awoonor-Williams, E.; Golosov, A. A.; Hornak, V. CovCysPredictor: Predicting Selective Covalently Modifiable Cysteines Using Protein Structure and Interpretable Machine Learning. *J. Chem. Inf. Model.* **2025**, *65*, 544–553.

(41) 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. *AAAS Research* **2022**, *2022*, 9873564.

(42) Abramson, J. et al. Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3. *Nature* **2024**, *630*, 493–500.

(43) Boatner, L. M.; Palafox, M. F.; Schweppe, D. K.; Backus, K. M. CysDB: A Human Cysteine Database Based on Experimental Quantitative Chemoproteomics. *Cell Chem. Biol.* **2023**, *30*, 683–698.e3.

(44) Boatner, L. M.; Eberhardt, J.; Shikwana, F.; Holcomb, M.; Lee, P.; Houk, K. N.; Forli, S.; Backus, K. M. CIAA: Integrated Proteomics and Structural Modeling for Understanding Cysteine Reactivity with Iodoacetamide Alkyne. *ACS Chem. Biol.* **2025**, *20*, 1669–1682.

(45) ESM Team. ESM Cambrian: Revealing the Mysteries of Proteins with Unsupervised Learning. https://www.evolutionaryscale.ai/blog/esm-cambrian.

(46) Uhlen, M.; Oksvold, P.; Fagerberg, L.; Lundberg, E.; Jonasson, K.; Forsberg, M.; Zwahlen, M.; Kampf, C.; Wester, K.; Hober, S.; Wernerus, H.; Björling, L.; Ponten, F. Towards a Knowledge-Based Human Protein Atlas. *Nat. Biotechnol.* **2010**, *28*, 1248–1250.

(47) Shen, M.; Dayhoff, G. W.; Shen, J. Protein Electrostatic Properties Are Fine-Tuned Through Evolution. 2025.

(48) Hayes, T. et al. Simulating 500 Million Years of Evolution with a Language Model. *Science* **2025**, *387*, 850–858.

(49) Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data. *Bioinformatics* **2012**, *28*, 3150–3152.

(50) Vig, J.; Madani, A.; Varshney, L. R.; Xiong, C.; Socher, R.; Rajani, N. F. BERTology Meets Biology: Interpreting Attention in Protein Language Models. ICLR 2021. 2021.

(51) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; Rives, A. Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model. *Science* **2023**, *379*, 1123–1130.

(52) Zhang, C.; Zhang, X.; Freddolino, L.; Zhang, Y. BioLiP2: An Updated Structure Database for Biologically Relevant Ligand–Protein Interactions. *Nucl. Acids Res.* **2024**, *52*, D404–D412.

(53) Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: A Public Database for Medicinal Chemistry, Computational Chemistry and Systems Pharmacology. *Nucl. Acids Res.* **2016**, *44*, D1045–D1053.

(54) Krapp, L. F.; Abriata, L. A.; Cortés Rodriguez, F.; Dal Peraro, M. PeSTo: Parameter-Free Geometric Deep Learning for Accurate Prediction of Protein

Binding Interfaces. *Nat. Commun.* **2023**, *14*, 2175.

(55) Löwe, J.; Li, H.; Downing, K.; Nogales, E. Refined Structure of $\alpha\beta$-Tubulin at 3.5 Å Resolution 1 1Edited by I. A. Wilson. *J. Mol. Biol.* **2001**, *313*, 1045–1057.

(56) Dayhoff II, G. W.; Uversky, V. N. Rapid Prediction and Analysis of Protein Intrinsic Disorder. *Protein Sci.* **2022**, *31*, e4496.

(57) Shen, M.; Kortzak, D.; Ambrozak, S.; Bhatnagar, S.; Buchanan, I.; Liu, R.; Shen, J. KaMLs for Predicting Protein p K_a Values and Ionization States: Are Trees All You Need? *J. Chem. Theory Comput.* **2025**, *21*, 1446–1458.

(58) 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.

(59) Liu, R.; Verma, N.; Henderson, J. A.; Zhan, S.; Shen, J. Profiling MAP Kinase Cysteines for Targeted Covalent Inhibitor Design. *RSC Med. Chem.* **2022**, *13*, 54–63.