Supporting Information:

Illuminating the Druggable Human Proteome with an AI Protein Profiling Platform

Guy W. Dayhoff II,[†] Daniel Kortzak,[†,¶] Ruibin Liu,[†] Mingzhe Shen,[†]

Zhong-Yin Zhang,[‡] and Jana Shen[*,†]

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

‡ *Borch Department of Medicinal Chemistry and Molecular Pharmacology, Purdue University, West Lafayette, IN 47907, U.S.A.*

¶ *Joint first author*

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

Contents

Supplemental Methods

At the core of AiPP is a pLLM representation-based clustering and label propagation framework that integrates heterogeneous experimental datasets into coherent ML training sets. This approach enables principled consensus labeling across noisy sources while preserving residue-level biochemical context. Critically, clustering also prevents representation-level data leakage during model evaluation—avoiding performance inflation from latent similarity in embedding space. Below we first describe the curation and formatting of all datasets used to train, validate, and benchmark AiPP.

1. Building LigCysABPP, LC3D, and LigBind3D databases

To train and evaluate the AiPP platform, we curated three orthogonal databases encompassing both covalent and reversible ligandability signals. These include LigCysABPP: cysteine ligandability dataset derived from the peptide-level cysteine reactivity profiles from 15 activity-based protein profiling (ABPP) studies published between 2016 and 2025;[S1–S15] LigCys3D*: a set of proteins with covalently liganded cysteines as captured by co-crystal structures derived from our previous larger database LigCys3D;[S16] and (iii) LigBind3D: ligand-binding residue annotations derived from the published database BioLip2.[S17]

Definition of records in the databases. All databases curated in this work were ultimately converted into a unified set of structured, site-level records, facilitating consistent downstream processing across modules. Each record captures residue-level biochemical annotations and includes the following fields: (1) a protein-specific unique identifier (`UID`), (2) a residue-of-interest (`ROI`) corresponding to a 1-based sequence position, (3) a binary label indicating ligandability (`EXP_BIN`), (4) source-specific quantitative metadata such as the raw competition ratio (`R`) in the ABPP experiment and the threshold (`EXP_THR`)

defined in the source study, (5) a source identifier (SOURCE), and (6) an optional NOTE field containing free-text metadata (e.g., ligand identity, molecular weight, probe identity). All records conform to this schema regardless of their original source— whether proteomic (LigCysABPP) or structural (LigCys3D and LigBind3D). Certain fields not applicable to all datasets—such as R and EXP_THR—were populated with pre-set values to preserve format consistency. For structure-derived records (LigCys3D and LigBind3D), R was set to 999 and EXP_THR to 0 to indicate direct structural evidence. For UNQUANTIFIED records, R was set to -1 and EXP_THR to 0 to denote inferred but unlabeled entries.

1.1 The LigCysABPP database

Curation of an ABPP dataset. We manually curated raw experimental data from 15 cysteine-directed activity-based protein profiling (ABPP) studies published between 2016 and 2025.[S1–S15] ABPP data reflect indirect chemoproteomic measurements of cysteine ligandability at scale.

For each ABPP study, we extracted the peptide-level entries and converted them into the structured, site-level records described above. Entries referencing multiple cysteine positions (ambiguous entries) or multiple UIDs (partially resolved, multi-value entries) were expanded into site-level (UID, ROI) records. Ambiguous entries were split into one record per ROI. In the case of multiple UIDs associated with a single ROI, multiple records were created, with the same ROI assigned to each UID. In the case of multiple UIDs associated with multiple ROIs, each entry was split into multiple records, each containing one UID and one ROI. Each derived record retained all original metadata and was annotated accordingly. Following the above treatment of the ambiguous and unresolved entries, we obtained a total of 683,192 records, spanning 58,704 unique cysteine sites (UID, ROI pairs) across 14,417 distinct proteins (UIDs).

Protein sequence reconstruction and record validation. To ensure consistency between input sequences and experimental annotations, all site-level records were subject to source-specific sequence reconstruction and cysteine identity validation. Each UID was mapped to its full-length canonical sequence using a local UniProt FASTA cache or the UniProt REST API (https://rest.uniprot.org). Records were excluded if the reported ROI did not correspond to a cysteine residue in the retrieved sequence, if the sequence contained non-canonical characters (e.g., 'X'), or if the accession was unavailable. In UniProt, the presence of "X" can indicate an unresolved internal region or, when present at the termini, that the entry is a fragment lacking part of the canonical sequence; both cases led to removal of the full record to ensure that only complete, biologically coherent proteins were used for modeling. The above protocol yield 671,515 validated records (out of 683,192 initial records), covering 53,867 unique cysteine sites across 12,745 proteins.

Consolidation of UIDs and filtering of derived records. To eliminate redundancy, we first collapsed UIDs referring to identical or subsequence protein entries, selecting a single representative UID per group. This reduced 650 UID groups in total (89 identical sequences and 561 subsequences), most of which corresponded to isoforms or redundant database entries. All records were then updated to reference the selected representative UID as the canonical identifier.

Next, we filtered multi-value–derived records: any record whose canonical UID did not appear in at least one unambiguous record (single UID, single ROI) was discarded, removing 1,607 records corresponding to 1,078 UIDs. All consolidation and filtering steps were logged to ensure reproducibility.

Treatment of unquantified cysteine sites in quantified proteins. We added 114,568 `UNQUANTIFIED` records, representing cysteine residues that were undetected by the ABPP probes but belonging to quantified proteins, i.e., those with at least one experimentally quantified cysteine site. These were added as negative records with $\texttt{EXP\_BIN} = \texttt{NEG}$ and

$\texttt{SOURCE} = \texttt{UNQUANTIFIED}$. These negative records are weakly labeled: they reflect the absence of observed reactivity rather than definitive evidence of non-ligandability, and were retained for provisional downstream consensus labeling (see section *Representation-Based Clustering and Label Derivation*).

Final records in the LigCysABPP database. Finally, we filter out records corresponding to proteins comprising fewer than 30 residues or more than 2046 residues (maximum ESM input size). This removed 81,341 records describing 25,485 unique cysteines across 1 small protein and 367 large proteins. The final ABPP database contained 703,135 records: 608,898 validated peptide-derived records and 94,237 UNQUANTIFIED records. Together, they span 140,459 unique cysteine sites (UID, ROI pairs), including 46,222 ABPP quantified and 94,237 unquantified (provisionally negative) cysteines across 10,649 proteins.

1.2 The LC3D database

LC3D dataset. LC3D was constructed as a LigCys3D-derived subset,[S16] providing a direct structural complement to the ABPP data. LigCys3D compiles cysteines observed to form covalent bonds in experimentally resolved structures from the RCSB Protein Data Bank (PDB). For LC3D, we retained only structures with covalent ligands of molecular weight ≥ 200 Da. Each structure was manually inspected, and entries deemed incorrect (e.g., wrong residue ID, no ligand, terminal cysteines) or unsuitable (e.g., cysteines forming disulfide bonds or helical staples, cysteines linked to peptides or proteins) were removed or corrected.

Sequence reconstruction and cysteine validation. Rather than relying on LigCys3D-provided annotations, we extracted the referenced PDB identifiers, downloaded the corresponding coordinate files, and independently derived the cysteine annotations. Se-

quences were reconstructed directly from atomic coordinates using our in-house tool (*pdb-doctor*), which integrates SEQRES records and REMARK 465 annotations to recover unresolved residues and ensure sequence–structure completeness. Site-level annotations were validated against these reconstructed sequences to confirm residue identity and positional accuracy. No remapping to UniProt was performed—in contrast to the LigCys3D database entries[S16]—as this could discard experimental details such as engineered mutations or truncations. Our aim was to preserve the exact protein sequence used in the structural experiment, rather than revert to a potentially mismatched canonical reference. After deduplication and validation, the LC3D database contained 316 POS LigCys3D-derived cysteine records across 275 unique proteins.

Provisional labeling of unliganded cysteines in co-crystal structures. To achieve complete cysteine coverage, we added the unlabeled cysteine residues present in the validated sequences as provisional negatives (EXP_BIN = NEG). As in LigCysABPP, these provisional negatives reflect the absence of observed covalent modification rather than definitive evidence, and were retained pending downstream consensus labeling (see *Representation-Based Clustering and Label Derivation*). After addition of these negatives, LC3D comprised 1,643 site-level cysteine records across 275 proteins. To ensure each protein contained both positive and negative sites, we filtered the set to retain only proteins with at least one negative cysteine. This removed 41 proteins, leaving LC3D with 234 proteins and 1,601 unique UID–ROI pairs: 274 positives (17.1%) and 1,327 negatives, with no masked sites.

Records in LC3D database. All entries were converted into the structured record format used for the LigCysABPP database. Each record used a UID in the format PDBID-CHAINID, with ROI values referencing sequence positions from the corresponding PDB structure. Ligand-specific metadata (e.g., molecular weight, three-letter PDB ligand code) were stored in the NOTE field. ABPP-specific fields such as EXP_THR and R were assigned preset

values (EXP_THR $= 0$, R $= 999$) to denote direct structural evidence.

1.3 The LigBind3D database

LigBind3D dataset. For developing models to predict reversible ligand-binding pockets, we independently curated an orthogonal structural dataset capturing non-covalent ligand-residue interactions from BioLiP2,[S17] a manually curated resource of biologically relevant protein-ligand complexes derived from the RCSB Protein Data Bank (PDB). We selected entries of monomer proteins containing small molecules with molecular weights between 150–600 Da, excluding those associated with nucleic acids, peptides, ions, crystallization agents, or cofactors.

For each selected PDB entry, we applied a custom structural annotation pipeline (PickPocket) to validate biologically relevant protein-ligand interactions. PickPocket resolves LINK records, merges "multi-residue" ligands, detects missing atoms, renumbers residues, and filters out artifacts. Ligands inferred by the RCSB but missing from coordinate files (e.g., covalently attached fragments) were reconciled when appropriate. Ligands were retained only if they (i) remained within the target molecular weight range after reconstruction, (ii) were not covalently linked to protein atoms—*except in cases where the covalent bond involved a cysteine residue*—and (iii) formed coherent binding pockets involving at least three spatially proximal residues.

Records in LigBind3D database. For each structure that passed all filtering steps, we reconstructed the full protein sequence directly from the atomic coordinate file, as in building the LC3D database. Next, ligand-contacting residues were labeled based on a 4.5 Å heavy-atom distance threshold. These residues were converted into structured, site-level records with EXP_BIN $=$ positive. Each record uses a UID in the format PDBID-CHAINID, with ROI values referencing sequence positions from the corresponding PDB structure. Ligand-specific metadata (e.g., 3-letter ligand code, molecular weight) were stored in the

NOTE field. As with the LC3D database, pre-set values (EXP_THR = 0, R = 999) were used to denote direct structural evidence. Residues not in contact with any ligand were added as negative records with EXP_BIN = NEG and are distinguishable via the SOURCE field. As in other datasets, these negative records are weakly labeled, reflecting the absence of observed interaction rather than confirmed non-binding, and were retained as provisional pending downstream consensus labeling (see *Representation-Based Clustering and Label Derivation*).

Overall, the final LigBind3D database comprised 686,255 site-level records across 1,998 unique proteins. This database offers a direct structural perspective on non-covalent ligand recognition, complementing the chemoproteomic data in ABPP and the covalent structural data in LC3D. However, unlike LigCys3D, which restricts ROIs to cysteines, LigBind3D treats every residue in each protein as a potential ROI, reflecting the broader diversity of reversible ligand interactions.

2. pLLM-based data clustering, label assignment and reconciliation

Per-token embeddings were extracted from layer 76 of ESM Cambrian (ESMC),[S18] a distinct 6-billion parameter protein language model. These 2,560-dimensional embeddings are derived directly from sequence and encode contextual, structural, and evolutionary features without requiring structural input. ESMC embeddings served as the unified representation for downstream clustering, label propagation, and supervised learning across both LigCys and LigBind modules. These unified embeddings provided a high-dimensional, functionally rich representation space from which we performed unsupervised clustering and downstream consensus-based label assignment.

Representation-based clustering. To group biochemically similar cysteine sites and prevent representation-level data leakage (see *Representation-Based Data Leakage Prevention*), we clustered validated site-level records using their ESMC-derived embeddings.

Each unique cysteine site—defined by a (UID, ROI) pair—was represented as a single node in embedding space, encoded by a 2,560-dimensional vector. All records corresponding to the same site were collapsed into a single node for clustering, ensuring that redundant evidence did not influence cluster formation or representative selection.

We computed a composite similarity score S between embedding vectors, defined as $S = \frac{1}{3}\big(C + D_1 + D_2\big)$, where C is the cosine similarity between vectors, and D_1 and D_2 are the inverse L1 and inverse L2 distances: $D_1 = 1/(1 + L_1)$, $D_2 = 1/(1 + L_2)$. Unlike cosine similarity alone, which captures only angular alignment and ignores vector magnitude, this composite score attains a maximum of 1 *only* when two embeddings are identical in both direction and length. This ensures that sites must be highly similar in full representation space to exceed the clustering threshold of $S > 0.3$, which approximately corresponds to a cosine similarity of 0.9. Empirically, we observed that this threshold effectively separates known functional neighbors while avoiding over-clustering of unrelated residues, and corresponds well to manually validated clusters in embedding space.

An undirected edge was drawn between two nodes if $S \geq 0.3$, and clusters were formed as connected components using a union–find algorithm. Within each cluster, a representative site was selected as the node with the highest average similarity S to all other members. Any node with $S < 0.3$ relative to the representative was pruned. Pruned nodes were iteratively re-clustered using the same procedure until convergence. Any remaining unclustered nodes were assigned as singleton clusters.

After clustering, all original records associated with each node were fully re-expanded so that each cluster included the complete set of source-level records corresponding to its constituent sites. This ensured that downstream consensus-based label assignment could integrate all available evidence during cluster-level label derivation. The final result was a set of compact, non-overlapping clusters of cysteine sites sharing high representation-level similarity, each anchored by a representative node used for consensus labeling and model training.

Label assignment for cysteine clusters. To denoise the cysteine ligandability labels from ABPP and LC3D databases, we performed cluster-level label assignment on the full set of representation-based clusters. To avoid data leakage, clustering was performed jointly on all validated ABPP and LC3D site-level records ($N = 704{,}730$), yielding 95,877 initial clusters (include 52,010 singletons) encompassing 141,898 unique cysteine sites.

LC3D records were treated as ground truth. Any cluster containing at least one LC3D record labeled `positive` was assigned a positive group label, irrespective of the number or source of other supporting records. Conversely, if a cluster contained LC3D records but none labeled `positive`, the group was assigned a negative label. In these cases, ABPP records were disregarded to preserve the orthogonality of LC3D labels. This scheme allowed direct structural evidence of covalent modification to propagate to biochemically similar cysteine sites not directly observed in structure but linked through high similarity in the representation space.

Clusters that did not contain any LC3D records were evaluated using a consensus voting scheme, in which positive or negative labels were assigned only if they met predefined thresholds across both record counts and sources (e.g., 1S–1R, 1S–2R, . . . , 4S–4R, 4S–5R, . . . ; see main text Table 1). For positives, a cluster was labeled positive once the required threshold was met (e.g., at the 1S–2R level, at least two `POS` records from one source; at the 4S–4R level, at least four `POS` records from different sources). Negatives followed the same thresholds but required the stricter condition that *all* records in the cluster were labeled `NEG` (i.e., no positive votes). Clusters that did not meet these criteria—or exhibited conflicting or insufficient evidence—were masked and excluded from training. Applying this rule, for example, at the 4S–4R level, 940 clusters were labeled `POS` and 5,170 clusters labeled `NEG`, covering 162,809 site-level records describing 11,473 distinct cysteines across 11,118 unique proteins. The remaining clusters, including 52,010 singletons, were masked.

Pruning LC3D-containing clusters to prevent data leakage. Following cluster-label assignment, we applied a conservative pruning step to eliminate potential data leakage. Specifically, all clusters containing any LC3D records were removed from the training pool after label assignment. Additionally, we identified all UIDs represented in these clusters and removed any remaining clusters that shared a UID with them—even if those clusters did not contain LC3D records themselves. This strict pruning ensured that no ABPP-derived cluster contained cysteine sites from proteins already seen in the LC3D-resolved set, eliminating indirect leakage via shared sequence context. Importantly, LC3D was withheld as an orthogonal benchmark set due to its high-confidence structural origin and minimal redundancy with ABPP-derived data. Excluding all LC3D-associated proteins from training ensures that subsequent model evaluation on this dataset reflects true generalization to unseen, structurally validated ligandable sites.

The remaining labeled clusters comprised a distilled, non-redundant training set derived exclusively from ABPP records. This set preserved biochemical and experimental diversity while enforcing rigorous separation between structurally derived and sequence-derived labels. All records within each retained cluster inherited the resolved cluster label, providing consistent and leakage-free supervision for model training.

Pruning an ABPP hold-out set. We created an ABPP hold-out set to provide a second, independent check during dataset expansion. As defined in the main text, it contains 23 proteins that were each measured by at least 10 independent ABPP sources and that have at least one positive and one negative cysteine. All site-level records for these proteins were flagged as hold-out and never used for model training or validation. In all steps (clustering, label assignment, and expansion), any cluster that contained a record from a hold-out protein—and any other cluster sharing that protein identifier (UID)—was placed entirely in the hold-out partition.

We used this hold-out set in two ways: (i) to monitor the LC3D-guided expansion for

signs of overfitting to LC3D, and (ii) to evaluate how the ABPP-guided cross-validation performs on unseen, well-measured proteins. Because cysteines within these proteins have different numbers of supporting measurements, we further stratified the hold-out cysteines into nine nS–nR subsets ($n = 1, \ldots, 9$) to examine how heterogeneous labeling affects fair performance assessment and error analysis. These diagnostics were purely observational and did not influence batch acceptance, hyperparameter choices, or model selection. A complete list of hold-out UIDs and cysteine records is provided in Supplemental Data (SD18–SD19).

Label assignment for LigBind clusters. LigBind3D data were processed independently from ABPP and LC3D datasets to derive cluster-level labels for residues involved in non-covalent ligand recognition. All LigBind3D records (686,255 site-level records across 1,998 unique proteins) were clustered using the same embedding-based similarity metric and threshold as in the LigCysABPP workflow.

Clusters were resolved using a permissive propagation rule: if any record in a cluster was labeled `POS` (i.e., within 4.5 Å of a ligand in a resolved structure), the entire cluster was assigned a positive label. Otherwise—if no records were labeled `POS`—the cluster was labeled `NEG`. No vote thresholds, source requirements, or masking logic were applied. This approach treats ligand proximity as direct structural evidence and enables propagation of contact labels to structurally or biochemically similar residues across isoforms, homologs, or convergent domains.

From the clustering step, we obtained 581,493 clusters (515,751 singletons; 65,742 multi-member) spanning 1,998 proteins and 687,712 LigBind3D records. Of these, 29,158 clusters were uniformly positive and 549,028 uniformly negative, while 3,307 contained mixed labels. Uniform clusters were defined as those in which all records shared the same label. At the record level, labels totaled 39,044 positive and 648,668 negative prior to reconciliation (5.7% positive; N:P = 16.6:1). To enforce within-cluster consis-

tency, we performed a single-pass reconciliation that converted 6,573 *False* $\rightarrow$ *True* assignments, yielding 45,617 positive and 642,095 negative records overall (6.6% positive; N:P = 14.1:1). These cluster-level labels constitute the high-confidence structural annotations used for LigBind model training and benchmarking.

Representation-based data leakage prevention. To ensure that model evaluation reflects true generalization, we used the representation-based clustering framework to control for data leakage. While conventional sequence identity filtering (e.g., via CD-HIT[S19]) can eliminate global sequence similarity, it fails to account for functionally convergent and/or structurally similar ligandable sites embedded within otherwise dissimilar proteins.

Our clustering operates on contextual ESMC embeddings, which capture local biochemical, structural, and evolutionary information. As a result, residues with similar ligandability profiles tend to cluster together—even across proteins lacking detectable sequence homology, yielding a higher-resolution view than traditional methods based on sequence alignment.

To prevent data leakage, we applied two complementary safeguards during dataset partitioning. First, each cluster was treated as an indivisible unit: no cluster was split across training, validation, or test sets. Second, we enforced protein-level (UID) exclusivity: all clusters containing records from the same protein (identified by UID) were assigned to the identical partition. This constraint was applied transitively—if any two clusters shared a UID (even through different residues), both were placed in the same partition. As a result, no protein appears in more than one dataset partition, and no structurally or biochemically similar residues are shared between training and evaluation sets. Together, these safeguards prevent both representation-level and protein-level leakage, ensuring that reported performance metrics reflect true predictive power on unseen proteins and local contexts.

3. LigCys and LigBind model architecture, training and validation

3.1 Sequence-Only models

LigCys-seq models. LigCys-seq models were trained to perform per-residue inference given the input sequence. The input protein sequence was tokenized and passed through ESMC,[S18] and per-token embeddings were extracted from layer 76, yielding a 2,560-dimensional vector for each cysteine. As the pretrained ESMC weights are not publicly available and the model cannot be fine-tuned, these embeddings were used in frozen form as input to a three-layer feedforward neural network with hidden dimensions of 1,024, 516, and 256, each followed by GELU activation. A layer normalization was applied to the input embeddings prior to the hidden stack, and a dropout layer ($p = 0.5$) was inserted before the final projection to a scalar logit representing the probability of cysteine ligandability.

Training is formulated as a binary classification task, with ground-truth labels of 1 (ligandable), 0 (non-ligandable), and 2 (masked). Masked residues are ignored during training and relabeled as 0 during evaluation. The AdamW optimizer is used with an initial learning rate of 1×10^{-5} and weight decay of 1×10^{-5}. Learning rate decay is applied multiplicatively at epochs 1, 2, and 4–9 by factors of 0.9, 0.7, and 0.5, respectively. Each model was trained for 10 epochs without early stopping. Twenty independent replicates were trained to support ensemble evaluation and assess variance.

Mini-batches were constructed per protein, initially containing all residues from a single input sequence. To address class imbalance, a 10:1 sampling ratio of negative to positive residues was enforced using a custom sampler. If a given protein contained too few residues to populate a full batch, additional residues were randomly sampled from other proteins in the training set to meet the batch size requirement. The loss function combined binary cross-entropy with focal modulation ($\alpha = 0.66$, $\gamma = 1.0$) and a positive-class weight of 0.1. Model checkpoints were selected based on maximum validation AUPRC and subsequently evaluated on held-out LC3D data using ensemble inference (see Section 3.3

Model evaluations based on ensemble inference).

LigBind-seq models. LigBind-seq models share the same backbone embedding procedure as LigCys-seq models, using fixed 2,560-dimensional per-token representations from ESMC layer 76. A simpler single-layer feedforward network was applied, consisting of a layer normalization, a GELU-activated linear transformation (2,560 $\rightarrow$ 2,560), dropout ($p = 0.1$), and a final projection to a scalar logit. This architecture was chosen based on empirical performance during preliminary screening and was sufficient to capture ligand-binding patterns from sequence alone.

Classification is binary, with residue-level labels of 1 (binding) or 0 (non-binding). No masked residues were present in the LigBind3D database. The loss function was adaptive focal loss with 20 confidence bins, initial $\gamma = 1.0$, bin updates every 5 epochs, a calibration-penalty term $\lambda = 0.1$, and γ clamped to [0.5, 5.0]. As with LigCys-seq models, all ESMC embeddings were frozen. Models were trained for up to 100 epochs using AdamW (learning rate 1×10^{-5}, weight decay 1×10^{-5}), with early stopping triggered after 20 epochs of no improvement in validation loss. For each of 10 train/validation splits, 20 models with different random seeds were trained, yielding a total of 200 LigBind-seq models.

Training batches were constructed per protein, enforcing a 15:1 ratio of negative to positive residues using the same sampling strategy as for LigCys-seq models. If a single protein did not contain enough residues to meet the minimum batch size of 256, additional samples were drawn randomly from other proteins. Evaluation metrics were computed on ensemble predictions (see *Ensemble Inference and Model Selection*) and included per-protein AUROC, AUPRC, F1 score, and top-K recovery.

3.2 Structure-Aware (SA) Models

Sequence-only models rely on contextual embeddings from ESMC. To incorporate explicit structural information, we developed structure-aware (SA) extensions. For LigCys, we introduced a *disjoint adaptive gating unit (DAGU)* to balance sequence embeddings with engineered structural features. For LigBind, we concatenated geometric embeddings with ESMC vectors without gating in this submission.

LigCys-SA models. LigCys-SA extends LigCys-seq by adding residue-level structural features computed from ESM3-predicted tertiary coordinates. Predicted PDBs were converted to solvent-excluded surface meshes with MSMS[S20] (probe radius 1.4 Å). For each residue we computed:

- **Solvent accessibility and topology:** per-atom SASA (freesasa[S21]) normalized by maximal residue SASA; centroid–surface distance; mesh curvature; surface patch area and concavity over a 5 Å geodesic neighborhood.
- **Binding-site proximity:** Euclidean and geodesic distances to ligand-binding residues predicted by LigBind-seq (encoded in B-factors), summarized within 3 Å and 6 Å (minimum, mean, median, standard deviation, counts).
- **Structural confidence:** per-residue pLDDT from the structure predictor.
- **Local chemical environment:** KaML-predicted absolute pK_a shifts; normalized SASA within 3 Å and 6 Å; ANCHOR2, VSL2B, and MDP disorder propensities from RIDA, each with radial statistics and counts.

Features were clipped to finite bounds, $\log_{1p}$-transformed where appropriate, and missing values were imputed to zero, yielding $\sim$70 raw features. To reduce redundancy, we removed near-zero-variance features and pruned one of each highly correlated pair

($|r| > 0.9$) after z-score normalization. The final non-redundant set comprised 45 features, concatenated with frozen 2,560-dimensional ESMC embeddings for training.

To integrate modalities, LigCys-SA used the DAGU block. At each residue, the input was split into an ESM block and a feature block. Each block was independently layer-normalized and gated by per-sample, channel-wise masks derived from mean-pooled block summaries passed through a linear layer with sigmoid activation. The gated blocks were concatenated, jointly layer-normalized, and passed through an N-layer MLP with GELU activations and dropout to produce a scalar logit. Optionally, the gating networks and block-wise normalization parameters were frozen after a user-defined epoch to stabilize blending.

Structural feature selection. The 45 selected features span six categories. Definitions and units are provided in Table S1:

- **Sequence–structure alignment:** `pKa_shift`, `norm_posn`, `plddt`.
- **Surface topology:** `surf_patch_area`, `surf_patch_concave_frac`.
- **Cysteine exposome:** `d_sg_exposure_depth`, `dir_sg_dist`, `mean_hydrophobicity_8A`, `density_net_charge_8A`, and class densities over 0–3 Å, 3–6 Å, and 6–9 Å shells.
- **Binding proximity:** `bind_euc_std`, `bind_geo_mean`, `bind_count_6A`.
- **KaML summary:** `rad_abs_shift_mean_3A`, `radial_avg_norm_sasa_3A`, `rad_abs_shift_mean_6A`.
- **RIDA summary:** `anchor_geo_min`, `anchor_geo_mean`, `vsl2b_label`, `vsl2b_euc_min`, `vsl2b_geo_min`, `vsl2b_count_6A`, `mdp_label`, `mdp_euc_min`, `mdp_euc_med`, `mdp_euc_std`, `mdp_geo_min`, `mdp_count_3A`.

Table S1: Non-redundant set of 45 residue-level structural features used in LigCys–SA.

Feature name	Definition
Sequence–structure alignment	
pKa_shift	KaML residue pK_a deviation
norm_posn	Normalized sequence position along chain (0–1)
plddt	Per-residue confidence score (0–100)
Surface topology	
surf_patch_area	Mesh area in 5 Å geodesic patch
surf_patch_concave_frac	Fraction of patch vertices with negative curvature
Cysteine exposome	
d_sg_exposure_depth	Offset of SG atom to nearest surface vertex
dir_sg_dist	Ray-traced outward SG distance
mean_hydrophobicity_8A	Mean Kyte–Doolittle scale within 8 Å
density_net_charge_8A	Net formal charge per residue within 8 Å
bin1.density.class	Fraction of residues by class, 0–3 Å shell
bin2.density.class	Fraction of residues by class, 3–6 Å shell
bin3.density.class	Fraction of residues by class, 6–9 Å shell
Binding proximity	
bind_euc_std	Std. of Euclidean distances to binding residues
bind_geo_mean	Mean geodesic distance to binding residues
bind_count_6A	Number of binding residues within 6 Å
KaML summaries	
rad_abs_shift_mean_3A	Neighbor average |pK_a shift|, 3 Å
radial_avg_norm_sasa_3A	Neighbor average normalized SASA, 3 Å

Continued on next page

Feature name	Definition
rad_abs_shift_mean_6A	Neighbor average |pK$_a$ shift|, 6 Å
RIDA summaries	
anchor_geo_min	Minimum geodesic distance to ANCHOR2 residues
anchor_geo_mean	Mean geodesic distance to ANCHOR2 residues
vsl2b_label	Binary indicator of VSL2B disorder region
vsl2b_euc_min	Minimum Euclidean distance to VSL2B residues
vsl2b_geo_min	Minimum geodesic distance to VSL2B residues
vsl2b_count_6A	Number of VSL2B residues within 6 Å
mdp_label	Binary indicator of MDP disorder region
mdp_euc_min	Minimum Euclidean distance to MDP residues
mdp_euc_med	Median Euclidean distance to MDP residues
mdp_euc_std	Std. of Euclidean distances to MDP residues
mdp_geo_min	Minimum geodesic distance to MDP residues
mdp_count_3A	Number of MDP residues within 3 Å

Features were computed from ESM3-predicted structures (MSMS surfaces; freesasa SASA), preprocessed by clipping to finite bounds, $\log_{1p}$ transforms where appropriate, zero-imputation for missing values, and z-score normalization for correlation screening. Redundancy was reduced by removing near-zero-variance features and pruning one of each pair with $|r| > 0.9$. Class "density" features denote fractions of residue classes within 0–3 Å, 3–6 Å, and 6–9 Å shells. Units: distances in Å, pLDDT on 0–100, charge unitless, counts as integers, pK$_a$ shifts in standard chemical units.

LigBind-SA models. LigBind-SA extends LigBind-seq by concatenating residue-level structural embeddings with ESMC representations. Structural embeddings were produced by an ensemble of eight pretrained geometric transformers (PeSTo architecture) applied to ESM3-predicted coordinates. Each transformer encoded one-hot atom types and 3D coordinates using a k-nearest-neighbor graph (up to 64 neighbors), updated rep-

resentations with attention layers that integrate distance and orientation features, and pooled atom-level vectors to residue-level embeddings via learned gating. Concatenated ensemble outputs were combined with frozen 2,560-dimensional ESMC vectors and fed to the same three-layer feedforward network as LigBind-seq (GELU activations, dropout $p = 0.1$). Unlike LigCys-SA, the DAGU block was not applied here. Training followed the LigBind-seq regimen (AdamW, learning rate 1×10^{-5}, weight decay 1×10^{-5}, batch size 256, up to 100 epochs with early stopping after 20 epochs without validation-loss improvement), with 20 replicates per split ensembled for final inference.

3.3 Ensemble inference

To enhance prediction stability and generalization performance, all models were evaluated using ensemble inference. For each task (LigCys and LigBind) and model variant (seq and SA), we constructed 10 non-overlapping data splits and trained 20 independent models per split, yielding 200 trained models per setting. All replicas used identical configurations but distinct random seeds and training shuffles.

To form an ensemble, we selected a single checkpoint from each replica corresponding to the epoch with highest validation AUPRC. These 200 best-performing models—one per replica—were then used to generate per-residue predictions for evaluation and ranked recovery (see *Model Evaluation and Ranked Recovery*). For LigCys models, we applied a top-k voting strategy with $k = 1$, whereas for LigBind models we employed probability averaging. Ensemble predictions were thus aggregated using either (i) average probability across models or (ii) per-protein top-k voting, in which each model contributed votes for its k highest-confidence residues within a given protein.

Unless otherwise stated, top-1 voting was used for all ensemble evaluations. In this setting, each model identifies the single residue in a protein with the highest predicted ligandability probability. A residue is labeled positive if it receives a majority of votes (i.e., from at least 101 of 200 models). This aggregation strategy reflects practical constraints

in experimental validation, where only a small number of high-confidence sites can be tested.

All metrics—including AUROC, AUPRC, F1 score, and top-k recovery—were computed on held-out test folds using ensemble outputs. Notably, the top-1 voting scheme induces a natural binary decision threshold of 0.5: residues that receive majority support (i.e., predicted as top-1 by at least 101 of 200 models) are labeled positive, while all others are labeled negative. This provides a consistent framework for computing both ranked recovery and threshold-based classification metrics from the same ensemble predictions. The overall strategy consistently outperformed individual replicas and reduced variance across splits, yielding robust, reproducible performance estimates.

Model evaluation and ranked recovery. All models were evaluated on held-out benchmark sets using standard classification metrics, including the threshold-independent metrics, area under the receiver operating characteristic curve (AUROC) and area under the precision–recall curve (AUPRC). To provide an interpretable breakdown of prediction behavior, we also computed confusion matrices at a fixed decision threshold of 0.5, reporting the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). From these, accuracy was computed as $(\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN})$, precision as $\mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$, recall (sensitivity) as $\mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$, specificity as $\mathrm{TN}/(\mathrm{TN} + \mathrm{FP})$, and F1 score as the harmonic mean of precision and recall: $2 \cdot (\mathrm{Precision} \cdot \mathrm{Recall})/(\mathrm{Precision} + \mathrm{Recall})$.

We also evaluated ranked retrieval performance using top-k recovery. For each test protein, residues were ranked by predicted ligandability probability. Given a set of N known proteins and ligandable residues L_i for protein $i, 1 \leq i \leq N$, and the set of residues with top-k predictions $R_{k,i}, 1 \leq i \leq N$, top-k recovery was defined as the fraction of true positive predictions in all $R_{k,i}$ and the maximum number of true positives that could be recovered at the given k threshold.

$$\text{Top-}k \text{ Recovery} = \frac{1}{M_k} \sum_{i=1}^{N} |L_i \cap R_{k,i}| \,,$$

where

$$M_k = \sum_{i=1}^{N} \min\{k, |L_i|\}$$

is above mentioned maximum of recoverable positives.

We emphasize Top-1 recovery (i.e., $k = 1$) the primary evaluation metric, reflecting the practical constraint in experimental screening where only a single site can be tested. This metric directly addresses the critical question: when limited to testing one predicted site, how accurate is the model's top recommendation?

Because negative labels in LC3D testset are provisional – reflecting the absence of observed interaction rather than confirmed inactivity – ranked recovery provides a more robust and biologically grounded assessment of model performance. This strategy accounts for the uncertainty inherent in weakly supervised data and aligns with the platform's goal of guiding residue prioritization for experimental validation. Accordingly, top-1 recovery is reported as the primary evaluation criterion across all benchmark comparisons.

All evaluation metrics were computed at the per-protein level, i.e., calculated separately for each protein and then averaged. This formulation is more stringent than global aggregation, as it mitigates bias from protein length or site count and emphasizes consistency across targets. Reporting exclusively per-protein metrics therefore provides a more realistic estimate of the performance an end user can expect when applying the model to individual proteins in proteome-wide screens.

3.4 Auxiliary Modules

RIDA module. The Rapid Intrinsic Disorder Analysis (RIDA) module provides per-residue annotations of intrinsic disorder and molecular recognition features (MoRFs) using the RIDAO engine,[S22] which integrates widely used disorder predictors into a unified interface.

For each input sequence, RIDAO reports per-residue predictions from all constituent tools and aggregates them into a mean disorder profile that includes ANCHOR2[S23] predictions of molecular recognition feature (MoRF) propensity. MoRFs are short disordered segments that undergo binding-induced folding.

Within AiPP, RIDA outputs are incorporated as hand-engineered structural features for the LigCys-SA model. These disorder-derived annotations provide local context relevant to residue accessibility and conformational flexibility, especially in partially structured or cryptically disordered regions. RIDA is not used in supervision, representation learning, or label derivation, and is applied uniformly to all sequences post hoc. By integrating disorder metrics directly into the LigCys-SA input feature set, AiPP leverages structural plasticity signals to aid ligandability predictions in dynamic or non-globular regions.

KaML-ESM. KaML-ESM[S24] is a sequence-based model for residue-level pK_a prediction, built from ESM embeddings and trained on curated experimental and synthetic datasets. It achieves highly accurate pK_a prediction across five types of ionizable residues (Asp, Glu, His, Cys, Lys). A particular strength of KaML-ESM is the superior performance in predicting cysteines that are deprotonated or titrating at physiological pH.

Within AiPP, KaML-ESM is used without modification to generate residue-level pKa values, which are incorporated as electrostatic features in the structure-aware LigCys-SA model. These values reflect local chemical environments and support identification of reactive cysteines. KaML-ESM is not used in supervision, label derivation, or clustering; its outputs are used solely as auxiliary inputs to enhance ligandability predictions without requiring physics-based modeling.

4. LigCys training data expansion

4.1 LC3D-guided data expansion

To increase the training data size and coverage while preserving high-confidence supervision, we applied an iterative data expansion procedure guided by model performance on the Top-1 recovery of LC3D cysteines. The protocol started from the 4S–4R dataset containing cluster representatives from proteins with both positive and negative cysteines.

The candidate pool. The candidate pool included cysteines meeting the same 4S–4R consensus threshold, but from proteins containing exclusively positive cysteines, exclusively negative cysteines, and non-representative cysteines of the cluster.

Expansion protocol. In each iteration, 100 batches, each with 175 cysteines (except for 80 cysteines in iteration 1, 100 in iteration 2, 150 in iteration 3) are randomly sampled from the candidate pool. Each batch was randomly assigned to the training or validation set, while maintaining the same training/validation data ratio of 9:1. For each batch, 24 models, generated from 6 random splits and each with 4 models (different random seeds), were trained and selected based on AUPRC on the validation set. The 24 models were used to generate ensemble predictions (see section 3.3 Ensemble inference) for the LC3D test set, and the batch that produced the model with the highest Top-1 recovery on the LC3D cysteines was accepted. All cysteines from the accepted batch, along with their labels, were then transferred from the candidate pool to the training pool. The iteration continues until Top-1 recovery no longer increases. The final expanded set (4S–4R-e) and the S10 hold-out test set were combined to train a production model, which was then evaluated on the LC3D dataset. The performance metrics of the production and models at each iteration are given in Table 2.

4.2 LigCysABPP-guided data expansion

An alternative to LC3D-guided data expansion is to systematically expand data based on cross validation AUPRC on the ABPP data. The baseline dataset from the LC3D-guided expansion (4S–4R, proteins with both positive and negative Cys, cluster representatives only) was used as the starting point for this alternative approach.

Batch construction. Candidate batches were assembled with a fixed composition of exactly 25 positives and the remainder negatives per batch. During the early iterations, when the training pool was relatively small, the batch size was set to 100 residues (25 positives and 75 negatives). As the pool expanded, the batch size was increased to 175 residues (25 positives and 150 negatives). Residues were drawn from the candidate set in order of *highest* ensemble uncertainty—i.e., increasing $|\,\overline{p} - 0.5\,|$ across baseline checkpoints—so that the most uncertain residues were evaluated earliest.

Cross-validation. We used group-aware 10-fold cross-validation, with three replicate models trained per fold using different random seeds, yielding a total of $K \times R = 30$ models for each baseline and candidate evaluation. The UID$\rightarrow$fold assignment was generated once at the start of the tournament, holding out 20% of UIDs for validation and requiring each validation subset to match the global positive fraction within 1%. This mapping was cached and reused throughout the tournament to minimize variance from resampling. At the beginning of each iteration, a baseline ensemble was trained on the current training pool using these cached folds, and the resulting checkpoints provided (i) a direct reference for scoring candidate batches and (ii) per-residue ensemble uncertainty scores used to prioritize residues during batch construction.

Controlling variance. Several safeguards ensured that observed performance changes reflected the added residues rather than randomness. A single UID$\rightarrow$fold mapping was

generated at the start of each tournament and cached for all evaluations within that tournament; validation subsets were constrained to match the global positive fraction within 1%. After each acceptance step, any newly added UIDs were deterministically assigned to folds to maintain balance while preserving existing assignments. To reduce long-term partitioning bias, the tournament was periodically restarted with freshly generated folds (after iterations 4, 8, and 12). All models within a given iteration (baseline and candidates) were initialized from the same saved weight snapshot and trained with identical architectures, losses, and optimization schedules (see "Model Architectures, Training, and Optimization"). Finally, three random-seed replicas were trained per fold, and performance was summarized as the mean across all $K \times R$ fold–replicate scores, with a 95% confidence interval computed over those $K \times R$ values. Together, these design choices minimized variance from partitioning and training stochasticity, isolating true performance gains attributable to the added residues.

Tournament evaluation. For each candidate batch, a provisional augmented training pool was formed by unmasking the corresponding residues in the candidate pool, and its performance was evaluated on the same cached folds used for the iteration's baseline ensemble. This ensured that differences in outcome reflected the added residues rather than variability in data partitioning. Performance was measured as AUPRC enrichment, defined as $\Delta = \mathrm{AUPRC} - \mathrm{prevalence}$, which quantifies improvement over the random baseline expected at the observed class balance. This normalization prevented shifts in prevalence across iterations from spuriously inflating or deflating apparent performance. Each batch was scored by the mean across all $K \times R$ fold–replicate scores, with a 95% confidence interval computed over those values; batches were ranked by the lower CI bound, emphasizing statistical robustness rather than raw average performance.

Recombination and selection. Each tournament iteration comprised up to four rounds. In the first round, all candidate batches were scored directly against the baseline en-

semble. In subsequent rounds, a hybrid genetic algorithm plus successive halving strategy was applied: the top fraction of batches (20% in early iterations, 50% in later ones) was retained, while the remainder were discarded. Retained batches were then recombined through one-point crossover (swapping segments of residues between two parent batches) and mutation (randomly altering a small fraction of residues to maintain diversity), with an exploration fraction $\epsilon = 0.1$ controlling the mutation rate, to generate new candidates for the next round. The best-performing batch from the previous round was always preserved.

Acceptance criteria and pool updates. At the end of each tournament iteration, the *overall* best-performing batch was accepted if its lower CI bound or its mean enrichment exceeded the baseline ensemble. Accepted residues were simultaneously merged into the training pool—updating existing UIDs or creating new records as needed—and masked in the candidate pool to prevent reselection. For each iteration, a detailed acceptance log was recorded and both check-pointed and augmented training pools were saved; the process terminated when no candidate batch met the acceptance threshold.

Supplemental Figures

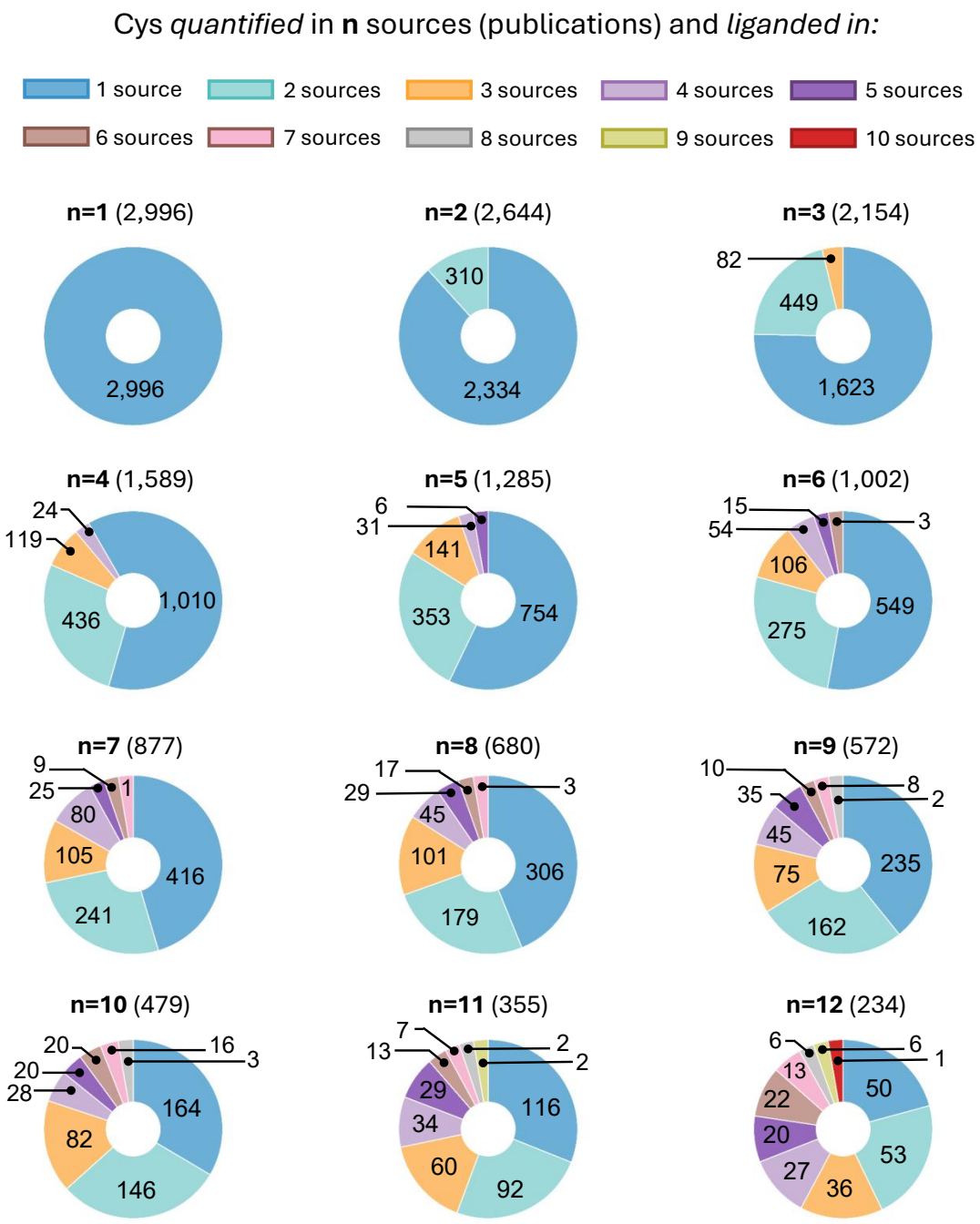

Figure S1: **Analysis of cysteine ligandability across ABPP sources and records.** Each pie chart displays the number of liganded cysteines quantified by n sources. A liganded cysteine is defined as one that has at least one pos ABPP record. Each pie segment represents the number of cysteines labeled pos in 1, 2,.. m sources. This visualization illustrates the level of consensus across different sources.

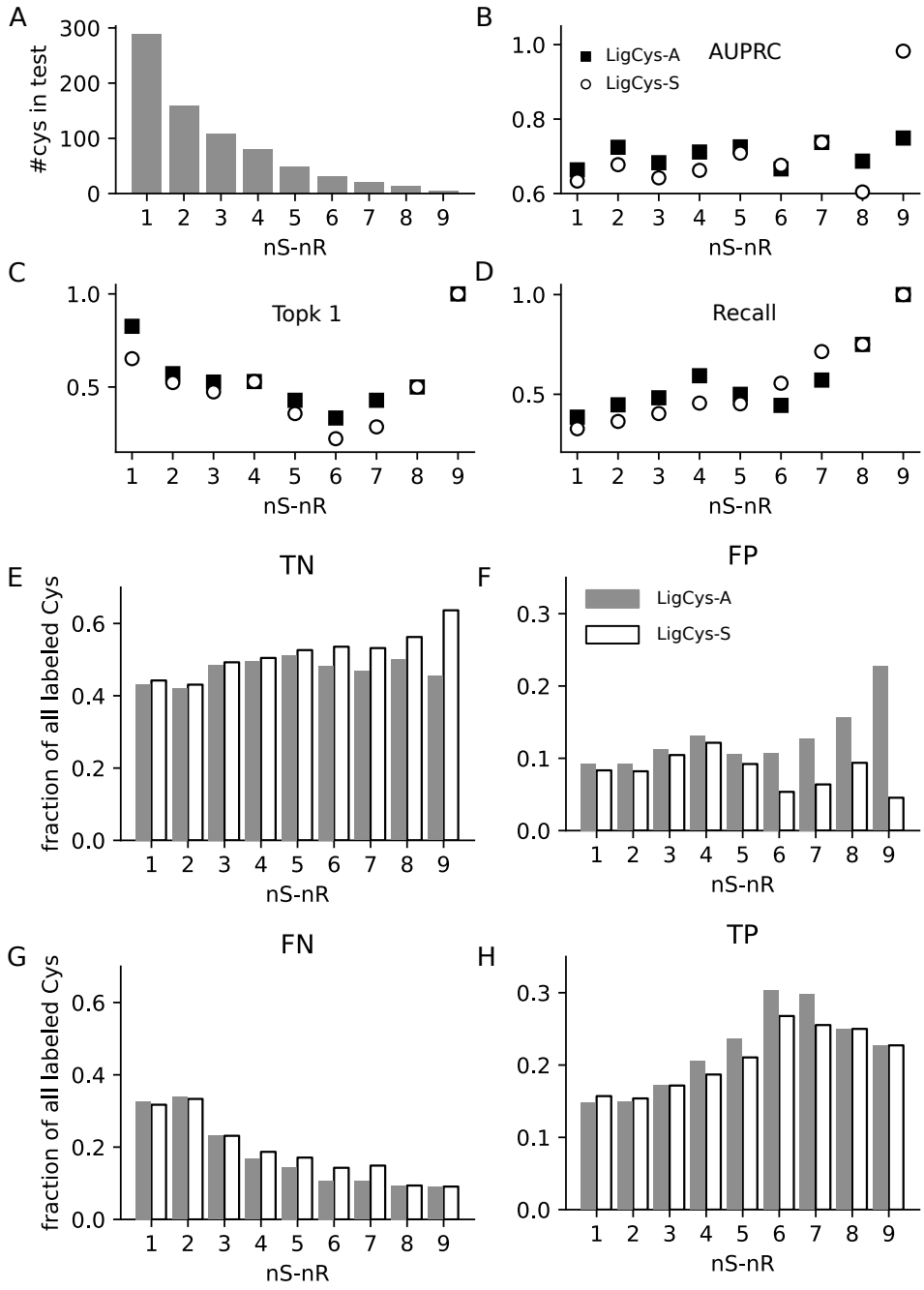

Figure S2: **Evaluation of LigCys-S and LigCys-A models on the ABPP hold-out set at different source consensus thresholds.** From the set of held-out proteins 9 test sets are derived based on nS-nR consensus thresholds (n=1,...,9). The LC3D-expanded (LigCys-S) and ABPP-expanded (LigCys-A) models are evaluated using these 9 test sets. **a.** Number of labeled cysteines in the nS-nR test sets. **b-d.** AUPRC, Top-1, and recall of LigCys-S and LigCys-A models in predicting liganded cysteines in the nS-nR test sets. **e-h.** Outcomes of the predictions by the LigCys-S (open bars) and LigCys-A (filled bars) models. Confusion matrix components, true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), are shown as proportions of the total labeled cysteines.

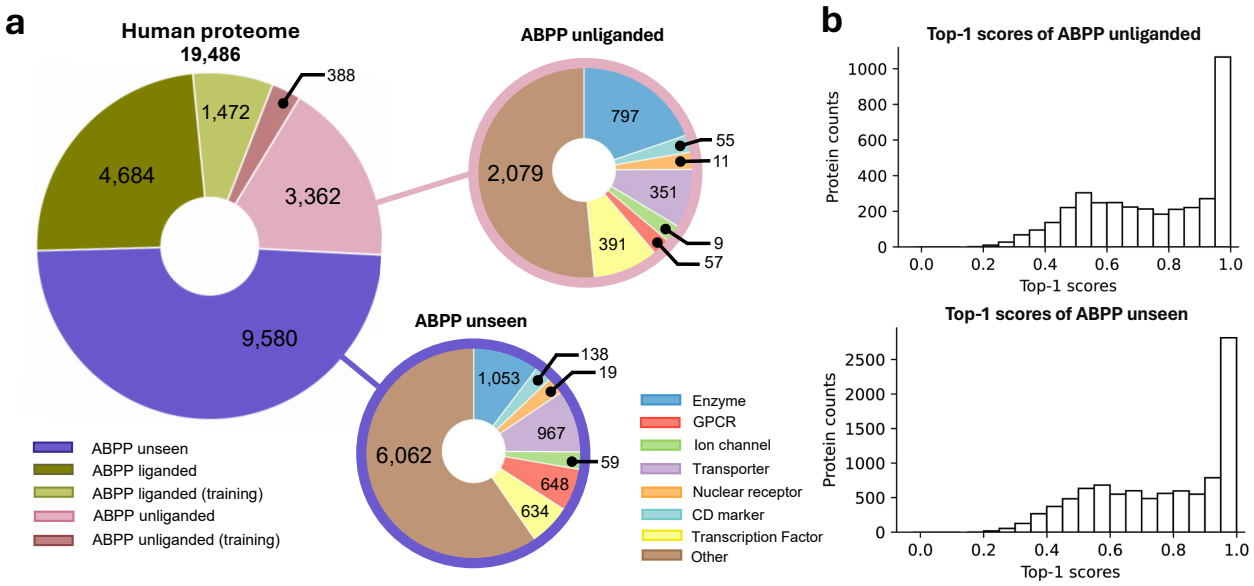

Figure S3: Functional family classification and LigCys Top-1 prediction scores for the ABPP unliganded or unseen proteins in the human proteome. This figure is to accompany Figure 7.

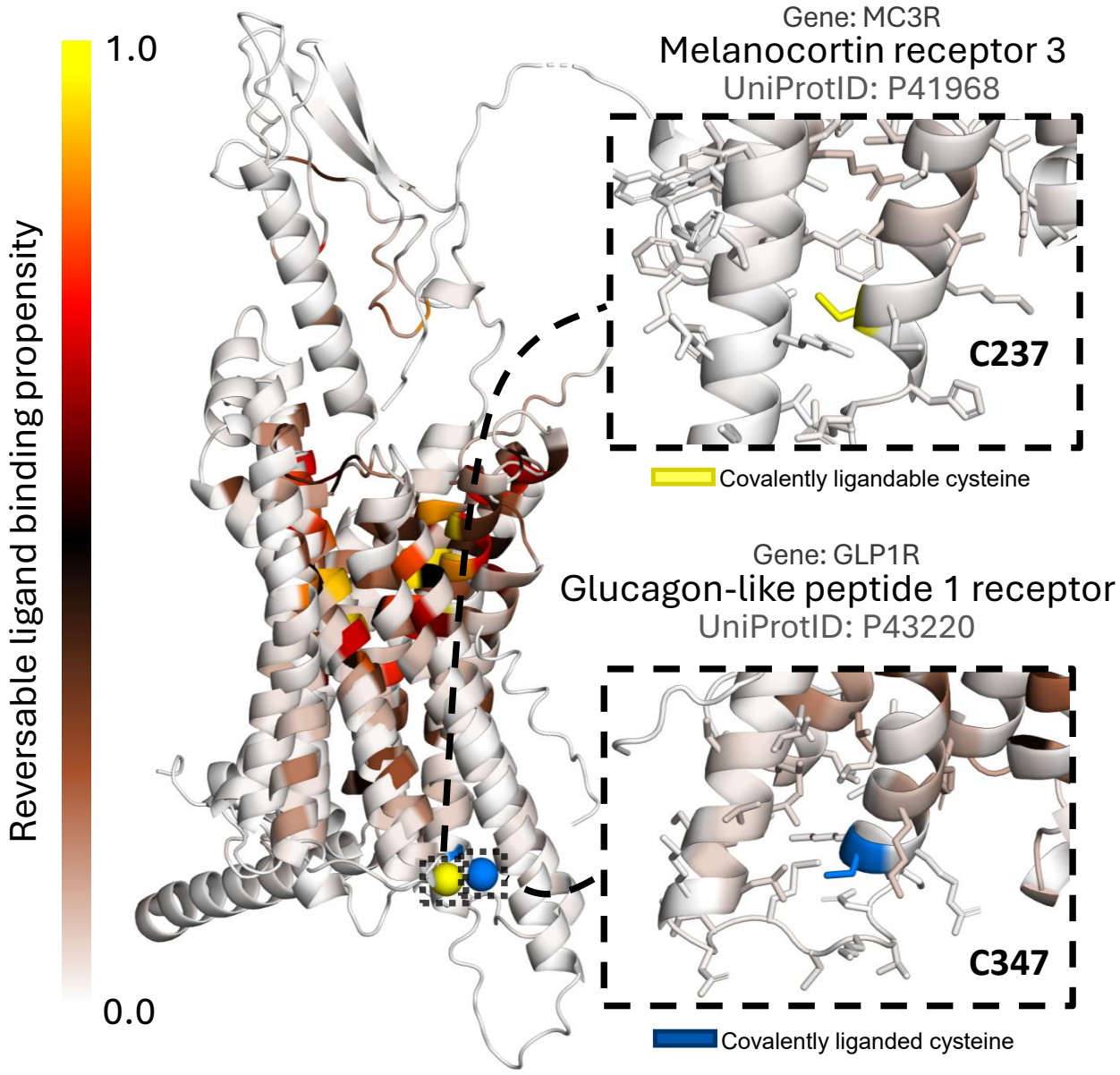

Figure S4: Reversible binding propensity scores for MC3R and GLP-1R mapped onto the ESM3 predicted (inactive) structures. The LigCys-predicted ligandable cysteine (C237$^{6.30}$) in MC3R and the analogous C347$^{6.31}$) in GLP-1R are colored yellow and blue, respectively. There nearby residues display low propensities (light pink color) for reversible binding.

References

(S1) 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.

(S2) Vinogradova, E. V.; Zhang, X.; Remillard, D.; Lazar, D. C.; Suciu, R. M.; Wang, Y.; Bianco, G.; Yamashita, Y.; Crowley, V. M.; Schafroth, M. A.; Yokoyama, M.; Konrad, D. B.; Lum, K. M.; Simon, G. M.; Kemper, E. K.; Lazear, M. R.; Yin, S.; Blewett, M. M.; Dix, M. M.; Nguyen, N.; Shokhirev, M. N.; Chin, E. N.; Lairson, L. L.; Melillo, B.; Schreiber, S. L.; Forli, S.; Teijaro, J. R.; Cravatt, B. F. An Activity-Guided Map of Electrophile-Cysteine Interactions in Primary Human T Cells. *Cell* **2020**, *182*, 1009–1026.e29.

(S3) 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.

(S4) 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.

(S5) 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.

(S6) 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.

(S7) 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.

(S8) 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.

(S9) 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.

(S10) 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.

(S11) Biggs, G. S.; Cawood, E. E.; Vuorinen, A.; McCarthy, W. J.; Wilders, H.; Riziotis, I. G.; Van Der Zouwen, A. J.; Pettinger, J.; Nightingale, L.; Chen, P.; Powell, A. J.; House, D.; Boulton, S. J.; Skehel, J. M.; Rittinger, K.; Bush, J. T. Robust Proteome Profiling of Cysteine-Reactive Fragments Using Label-Free Chemoproteomics. *Nat. Commun.* **2025**, *16*, 73.

(S12) 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.

(S13) 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. Chemical Proteomics Identifies Druggable Vulnerabilities in a Genetically Defined Cancer. *Cell* **2017**, *171*, 696–709.e23.

(S14) 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.

(S15) Burton, N. R.; Backus, K. M. Functionalizing Tandem Mass Tags for Streamlining Click-Based Quantitative Chemoproteomics. *Commun. Chem.* **2024**, *7*, 80.

(S16) 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.

(S17) 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.

(S18) ESM Team. ESM Cambrian: Revealing the Mysteries of Proteins with Unsupervised Learning. https://www.evolutionaryscale.ai/blog/esm-cambrian.

(S19) 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.

(S20) Sanner, M. F.; Olson, A. J.; Spehner, J.-C. Reduced Surface: An Efficient Way to Compute Molecular Surfaces. *Biopolymers* **1996**, *38*, 305–320.

(S21) Mitternacht, S. FreeSASA: An Open Source C Library for Solvent Accessible Surface Area Calculations. *F1000Res.* **2016**, *5*, 189.

(S22) Dayhoff II, G. W.; Uversky, V. N. Rapid Prediction and Analysis of Protein Intrinsic Disorder. *Protein Sci.* **2022**, *31*, e4496.

(S23) Mészáros, B.; Erdős, G.; Dosztányi, Z. IUPred2A: Context-Dependent Prediction of Protein Disorder as a Function of Redox State and Protein Binding. *Nucl. Acids Res.* **2018**, *46*, W329–W337.

(S24) Shen, M.; Dayhoff, G. W.; Shen, J. Protein Electrostatic Properties Are Fine-Tuned Through Evolution. 2025.