

CysDB: A Human Cysteine Database based on Experimental Quantitative Chemoproteomics

Lisa M. Boatner^{1,2}, Maria F. Palafox³, Devin K. Schweppe⁴ and Keriann M. Backus^{1,2,5,6,7,8*}

1. Biological Chemistry Department, David Geffen School of Medicine, UCLA, Los Angeles, CA, 90095, USA.
2. Department of Chemistry and Biochemistry, UCLA, Los Angeles, CA, 90095, USA.
3. Department of Human Genetics, David Geffen School of Medicine, UCLA, Los Angeles, CA, 90095, USA.
4. Department of Genome Sciences, University of Washington, Seattle, WA, 98185, USA.
5. Molecular Biology Institute, UCLA, Los Angeles, CA, 90095, USA.
6. DOE Institute for Genomics and Proteomics, UCLA, Los Angeles, CA, 90095, USA.
7. Jonsson Comprehensive Cancer Center, UCLA, Los Angeles, CA, 90095, USA.
8. Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, UCLA, Los Angeles, CA, 90095, USA.

*Corresponding Author: kbackus@mednet.ucla.edu

ABSTRACT

Cysteine chemoproteomics studies provide proteome-wide portraits of the ligandability or potential ‘druggability’ of thousands of cysteine residues. Consequently, these studies are enabling resources for closing the druggability gap, namely achieving pharmacological manipulation of ~96% of the human proteome that remains untargeted by FDA approved small molecules. Recent interactive dataset repositories, such as OxiMouse and SLCABPP, have enabled users to interface more readily with cysteine chemoproteomics studies^{1,2}. However, these databases remain limited to single studies and therefore do not provide a mechanism to perform cross-study analyses. Here we report CysDB as a curated community-wide repository of human cysteine chemoproteomics data that incorporates high coverage data derived from nine studies generated by the Backus, Cravatt, Gygi, Wang, and Yang research groups. CysDB is a SQL relational database that is publicly available at <https://backuslab.shinyapps.io/cysdb/> and features chemoproteomic measures of identification, hyperreactivity, and ligandability for 62,888 cysteines (24% of all cysteines the human proteome). The CysDB web application also includes annotations of functionality (UniProtKB/Swiss-Prot, Pfam, Panther), known druggability (FDA approved targets, DrugBank, ChEMBL), disease-relevance and genetic variation (ClinVar, Cancer Gene Census, Online Mendelian Inheritance in Man), and structural features (Protein Data Bank). Showcasing the utility of CysDB, here we report the discovery and enrichment of ligandable cysteines in undruggable classes of proteins, the observation that a subset of cysteines showed marked preference for specific classes of electrophiles (chloroacetamide vs acrylamide), and that ligandable cysteines are present in numerous undrugged disease-relevant proteins. Most importantly, we have designed CysDB for the incorporation of new datasets and features to support the continued growth of the druggable cysteinome.

INTRODUCTION

Small molecule chemical probes are useful tools for modulating protein function that can

serve as leads for future medications. Therefore, ongoing efforts in the chemical biology community have set ambitious goals in matching every protein with a chemical probe³. Complicating matters, <4% of the human proteome has been pharmacologically targeted by an FDA approved small molecule. Cysteine chemoproteomics has emerged as an enabling technology that addresses this druggability gap by identifying thousands of functional and potentially druggable cysteines proteome-wide¹⁻²⁵. Demonstrating this utility, prior cysteine chemoproteomic studies, including our own, have revealed a strikingly low overlap between proteins containing 'ligandable' or potentially 'druggable' cysteines and those that have been targeted by FDA approved molecules¹¹.

Cysteine proteomics experiments can be generally classified into four main categories: (1) identification, (2) measuring hyperreactivity, (3) measuring ligandability and (4) measuring redox state. (1) We consider identification studies as those aiming to increase coverage of cysteine containing peptides⁴⁻⁶. (2) Hyperreactivity experiments measure the intrinsic reactivity of cysteines towards highly electrophilic probes⁷⁻¹⁰, while (3) ligandability experiments measure the intrinsic ligandability or potential 'druggability' of cysteines using libraries of drug-like electrophilic molecules, natural products, and lipid derived electrophiles^{2,11,15-19}. (4) Finally, redox protocols are tailored to identify redox sensitive cysteines^{1,20-23}.

While the overarching objectives of these studies are non-redundant, they do share general features, including conceptually similar workflows and, most importantly, shared targets. In a standard cysteine chemoproteomics experiment for example, the proteome is treated with a pan-cysteine reactive probe, followed by enrichment on streptavidin resin, sequence specific proteolysis, and tandem liquid chromatography mass spectrometry analysis (LC-MS/MS).

Despite considerable recent advances in instrumentation, sample preparation, and data analysis, most cysteine chemoproteomics studies only sample a small fraction of all cysteines in the proteome, with the highest coverage studies sampling ~13% of all cysteines^{1,7,9}. Reasons for this gap include protein abundance and restricted expression profiles, location of cysteines in very

long or very short tryptic peptides, which are not detected in standard trypsin digests, and unreactive cysteines, such as those buried in the protein core or located in structural disulfides. Despite these technical limitations, the cysteinome continues to grow, with the addition of multiple high coverage new studies in 2022 alone^{6,10,14}.

The availability of easily searchable cysteine databases—including Oximouse¹, the Ligandable Cysteine Database, and previously reported Cysteinome²⁴—has increased the general accessibility of these large proteomics datasets, allowing rapid queries for targets of interest^{9,12,13}. However, with the exception of the Cysteinome database, which was launched in 2016 and is no longer publicly accessible, these databases are restricted to datasets derived from single publications.

To facilitate future studies aimed at global or target focused analyses of the cysteinome, we envisioned the establishment of a unified cysteine-focused database that would fulfill the following criteria. First, the database would incorporate datasets from many large scale cysteinomic studies and therefore enable rapid and facile inter- and intra-dataset comparisons. Second, the database would include information about the reactivity and ligandability of cysteines together with the druggability of their corresponding proteins, as indicated by availability of FDA approved drugs. Lastly, and most significantly, the database would integrate functional and structural data from the UniProtKB/Swiss-Prot, Cancer Gene Census (CGC), ClinVar, Human Protein Atlas (HPA), ChEMBL, DrugBank and the Protein Data Bank (PDB)²⁶⁻³², to enable prioritization of targets for future studies. Here we present the CysDB, which is an interactive database that fulfills these criteria for 62,888 cysteines and 11,621 proteins. Importantly, to promote the continued growth of cysteine chemoproteomics, we also provide a straightforward route for addition of future datasets.

RESULTS

(1) Data curation to establish a set of processed and aggregated chemoproteomics datasets to enable CysDB.

Our first step towards creating CysDB was to assemble a set of publicly available datasets. With the overarching goal of establishing a high coverage and highly curated database of human chemoproteomics studies to enable cross-dataset exploration, we opted to focus on a reduced set of available datasets. We prioritized studies that reported high coverage datasets that measured one or more of the following parameters: (1) total number of cysteines identifiable by the pan-cysteine reactive probes, (2) measurement of cysteine intrinsic reactivity towards iodoacetamide alkyne (**IAA, 1**; **Figure 1A** and **Figure S1**) and (3) assaying cysteine ligandability (**Figure 1A** and **Figure S2**). In total, we collected nine datasets that fulfilled our criteria (**Figure 1B** for all datasets used)^{2,4-11}.

Notably, all these studies rely on the same general cysteine chemoproteomic workflow: cells or lysates are treated with a cysteine reactive probe (**Figure 1A**, iodoacetamide alkyne (**IAA, 1**) or an iodoacetamide desthiobiotin reagent (e.g., **DBIA**² or **IA-DTB**⁸) to cap all accessible cysteines. Labeled proteins are subjected to enrichment on streptavidin or related resins together with sequence specific proteolysis followed by liquid chromatography-tandem mass spectrometry (LC-MS/MS). Several of our included studies⁷⁻⁹ further classify cysteine intrinsic reactivity and pinpoint hyperreactive cysteines by comparing relative cysteine labeling by two concentrations (10x and 1x) of cysteine enrichment handle (**Figure 1A** and **Figure S1**). Signal intensity differences between 100 μ M and 10 μ M treated proteomes are reflected by a ratio ($R_{[high]:[low]}$). Hyperreactive cysteines are defined as those with $R_{10:1}$ values < 2, indicating labeling events that are not concentration dependent. Most included studies provide a metric of cysteine ligandability or putative druggability^{2,4-5,8,10-11}, which is generated by comparing relative labeling by equimolar iodoacetamide in the presence and absence of electrophilic compound, with decreased labeling indicative of a high occupancy labeling event (**Figure 1A** and **Figure S2**).

To produce a rigorously curated database, we subjected all prioritized datasets to a series of data processing steps tailored to the nature of the study. First, we aggregated all non-redundant cysteines published by all studies, using the unique identifier UniProtKBID_CYS#. For some studies^{2,4-9,11} residue positions and protein identifiers were provided in the supporting information. For a subset of studies, the supporting tables instead provided labeled peptide sequences and protein IDs^{7,10}. To merge these two data types, we mapped each peptide to the corresponding canonical protein sequence using the UniProtKB reference FASTA from January 2022—this approach recovered nearly all cysteines, with only 37 dropped due to mismapping (**Table S1**), likely caused by differences in UniProtKB releases used in dataset search, as observed in our prior study⁹. In the event of proteomic analyses comparing cysteine labeling using different experimental conditions (e.g., unstimulated versus stimulated cells), we opted to incorporate only the datasets derived from control (no treatment) conditions, with the goal of limiting the potential impact of cell-state dependent differences of cysteine reactivity as a potential confounder to our downstream analyses. To address the many additional parameters, including data analysis pipeline differences, cysteines with incorrect residue numbers and peptides that match to multiple protein sequences (2,823 entries), we include the UniProtKB release and software used to process mass-spectrometry data for each dataset in **Table S1**^{7,18,33-38}. Aggregation of all datasets, including results from using multiple cell lines^{2,4-11}, resulted in the chemoproteomic identification of 62,888 unique cysteines and 11,621 proteins (**Figure 1C** and **1D**), which to our knowledge represents the most comprehensive cysteinome dataset reported to date.

Using the studies reporting measures of cysteine ligandability or labeling by electrophilic fragments or druglike molecules, we further stratified our dataset to generate a master set of all ligandable cysteines. The datasets included in our database (**Figure 1A**) were all prepared using the same general workflow where samples (lysates or cells) were treated by either a vehicle (DMSO) or a cysteine-reactive electrophile functionalized compound and the compound-dependent changes in IAA or IADB reactivity assayed by LC-MS/MS analysis. Prior analyses

have revealed that comparable competition ratios can be calculated using either MS1 or MS2 level quantification^{2,4-5,8,10-11}. Therefore, we opted not to differentiate between samples analyzed using different quantification methods, including isotopic labeling strategy (TMT or isotopically enriched biotinylation reagents)^{2,6}, label free quantification and data independent acquisition (DIA) based MS2 level quantification (**Figure S2** for general workflow)^{4,8,10}. The vast majority (97.2%) of all compounds screened were found to be functionalized with either a chloroacetamide or acrylamide moieties (**Figure S3**). A small but notable subset of compounds did however feature alternative electrophiles, including covalent reversible cyanoacrylamides³⁸, fumarates, and activated esters—while activated esters are primarily lysine reactive our prior data indicates that they do also exhibit cysteine-reactivity^{40,41}.

All datasets included in our database relied on competition ratio cutoffs for what defines a cysteine as 'ligandable.' Generally, cysteines were categorized as liganded if they had at least two ratios $R \geq 4$ (hit fragments) and one ratio between 0.5 and 2 (control fragments). However, when processing the ligandability data for each dataset, we observed manuscript-specific differences in either the ratio cutoff value or number of minimum unique hit fragments (1 or 2) required to have the associated ratio cutoff value for designating a cysteine as ligandable. For example, Cao et. al. 2021 implemented a slightly more permissive ratio cutoff of 3 to account for high field asymmetric waveform ion mobility spectrometry (FAIMS)-induced ratio-compression⁵. By comparison, Vinogradova et. al. 2020 implemented a more stringent ratio cutoff of 5⁸. Another case we encountered was the inclusion of 'ligandable' cysteines where the unique identifier contained multiple modified cysteine residues, such as UniProtKBID_CYS#1_C#2. These types of identifiers are derived from peptide sequences simultaneously labeled with capture reagents at multiple cysteine residues (C1*XXXC5*) within the same sequence. Based on our experience with such peptides yielding noisy ratios, we opted to remove them from CysDB—a total of 2,584 peptides were excluded due to this criteria. Otherwise, despite the differences in defining ligandability, we opted to retain all remaining liganded cysteines to accurately represent each

study's reported findings (criteria for ligandability applied to each study is available in **Table S1**). In aggregate across all ligandability studies, a total of 43,475 unique cysteines (**Table S2**) had quantified ratios, and 9,246 unique cysteines were deemed ligandable. These cysteines were found in 4,404 proteins (**Figure 1C** and **1D**).

Next, we parsed processed data from published datasets measuring cysteine hyperreactivity⁷⁻⁹. The three hyperreactivity studies included in CysDB measured the relative IAA reactivity towards two concentrations of IAA (100 μ M and 10 μ M), where a quantitative isoTOP-ABPP ratio ($R_{[high]:[low]}$) reflects the differences in signal intensities between the 100 μ M and 10 μ M treated proteomes. Highly reactive cysteines, termed 'hyperreactive' residues, are identified as those that exhibit saturation or near-saturation of labeling at the lower IAA concentration. All three publications utilized the same numerical ranges to delineate cysteines into 'high,' 'medium,' and 'low' reactivity subsets, with high reactivity, also termed 'hyperreactive' residues as those with an $R_{10:1} < 2$, medium reactive cysteines between $R_{100:10} \geq 2$ and $R_{10:1} < 5$ and low reactivity cysteines $R_{10:1} > 5$. During dataset processing, we observed that Weerapana et. al. 2010 and Palafox et. al. 2021 report median values of all the replicates for each individual measure of cysteine reactivity, as well as an overall mean of medians to quantify the average reactivity per cysteine. In contrast, Vinogradova et al. reports the average of medians across all measurements. To accommodate these dataset dependent differences, we opted to report the mean of median ratio values for each detected cysteine. In aggregate, 8,604 cysteines on 4,032 proteins were quantified by these three studies, which resulted in identification of 489 hyperreactive cysteines and 426 proteins containing hyperreactive cysteines (**Figure 1C** and **1D**).

Collectively across all cysteines identified through our data aggregation efforts, 14% were deemed ligandable and less than 1% determined to be hyperreactive. Cross-dataset comparisons reveal the highest overall coverage dataset was reported by Yan et. al 2021 (**Figure 1E** and **Figure S4**)⁴, where an optimized SP3-FAIMS strategy was applied to analyze the proteomes of seven cell lines, which in aggregate identified more than 34,000 cysteines on 9,714 proteins from

7 cell lines (**Figure S4** and **S5**). A key outcome of the dataset aggregation required to build CysDB is an effective doubling of the size of the identified cysteinome. Collectively across all studies analyzed in CysDB, ~25% of all cysteines found on 57% of human proteins in UniProtKB have been assayed at least once by chemoproteomics (**Figure 1C** and **1D**).

(2) Establishing an SQL database with an RShiny user interface for CysDB

With a complete, curated dataset in hand, we constructed the CysDB database and web user interface outlined in **Figure 2A**. Processed data from prioritized studies (Supplementary Tables listed in **Table S1**)^{2,4-11} were prepared into a standardized input format for SQL integration (See **Table S1** for example data format and required information for future data integration to CysDB) and loaded into a database hosted in Google Cloud using MySQL v.8.0. (See Methods for more details on data preparation and processing). CysDB is a relational database composed of six individual tables (**SI Figure S6**). For public accessibility of CysDB, we developed a front-end, user interface powered by the Shiny framework (**Figure 2B**). Shiny converts queries from remote users into visualizations and results that are displayed on a web browser. Not only does our web application access the Cloud CysDB, but it additionally calls from both structural and functional external databases, including UniProtKB, COSMIC, ClinVar and PDB^{26-29,32}.

One challenge we faced during our processing of the data provided in each study's Supplementary Tables, was one-to-one mapping of protein accessions to gene names for SQL querying. For gene-centric queries, not all HUGO Gene Nomenclature Committee (HGNC)⁴² or Entrez gene symbols are associated with a single protein. Gene sequences translated to the same protein sequence can lead to multi-mapping of various gene names to one UniProtKB accession⁹. In CysDB, we found that 16 UniProtKB entries were associated with multiple gene names (**Table S1**). To address this limitation, we included the capability to search from the main landing page using gene symbols, protein names or disease terms. Querying with any of these terms displays a table listing associated UniProtKB entries. The user then selects one of the listed

UniProtKB accessions for CysDB search. The CysDB RShiny interface enables the user to interact with cysteine chemoproteomics datasets, generate personalized figures, and download their results. Anywhere in the app, a user can save graphs as an image by clicking on a camera button at the top right corner and export query results to a CSV file by clicking a download button at the bottom of a table. The CysDB app includes five sections: Protein, Mutation, Enrichment, Compound, Statistics, and Datasets.

First, users can visualize the CysDB data in a protein-centric manner by selecting the protein explorer button, which is found on the home page (**Figure 3A**). Search for protein of interest (POI) by querying a UniProtKB ID returns the 'Protein Section,' which is further broken up into three separate tabs detailing activity, structure, and function. The activity tab provides a 'site map' indicating whether any cysteines in the POI are hyperreactive or ligandable together with the measured reactivity, measured competition ratios and the structures of all compounds that ligand the POI. The structure tab provides the user with annotations of proximal active site and binding site residues in both linear sequence and three-dimensional space and an easily accessible mechanism to visualize the three-dimensional protein microenvironment of chemoproteomic detected cysteines, including for structures reported in the PDB. Lastly, the function tab reports functional annotations for the POI generated from UniProtKB Gene Ontology (GO), and Reactome^{26,43,44}.

The 'Mutation Section' of CysDB, which can be accessed by selecting the 'Disease Explorer' button on the homepage, provides information complementary to that presented in the 'Protein Explorer' section. Query for a POI yields the aggregate number of CysDB cysteines, missense variants identified in ClinVar, the public repository of relationships between human genetic variation and phenotype, and cancer gene census (CGC) genes mapped to the POI. Search also generates a one-dimensional depiction of the corresponding protein sequence decorated with the positions of CysDB ligandable and hyperreactive cysteines alongside individual missense variants, sequence elements, and known ligand binding sites (**Figure 3B**).

To expedite identification of clinically relevant protein regions containing ligandable and hyperreactive cysteines, the Mutation Section of CysDB also provides the clinical significance for variants as reported by ClinVar²⁸, the public repository of relationships between human genetic variation and phenotype. To further enable pinpointing of cysteines relevant to human health, CysDB also provides CGC annotations of tumor types associated with POI, where relevant.

Looking beyond individual POIs, the ‘Enrichment Section’ of CysDB was built to enable facile visualization and analysis of the aggregated CysDB datasets. Global analyses provided include functional pathway, ontology, and disease enrichments of CysDB categories. By mapping the UniProtKB protein identifiers to Entrez gene symbols, CysDB also enables user-directed enrichment analysis of the ligandable and hyperreactive cysteine subsets, powered by the Enrichr package^{45,46} (**Figure 3C**).

As with the dataset-wide meta-analysis provided by Enrichment Section, the ‘Compound section’ of CysDB provides users with a global perspective of the electrophilic compounds employed in the CysDB cysteine ligandability studies. This portion of CysDB includes details of each molecule used in the ligandability experiments, including the publication name of each compound, corresponding CysDB names for each corresponding compound and dataset in an easily downloadable table. For CysDB, we created two naming conventions for each compound, a ‘Group Compound Identifier’ and an ‘Individual Compound Identifier. The Group Compound Identifier is called “GROUP_WARHEAD_#” and is based on only on SMILES strings; the individual compound identifier is called “WARHEAD_#,” and is unique for each SMILES string, cell line and publication author combination. With these two identifiers, one can group results from the same molecule generated under different conditions (e.g., sample preparation protocol, cell lines, etc.). Before generating these two compound identifiers, we had to standardize the format of each SMILES strings. Consistent with previous studies⁴⁷, we found that the molecular connectivity for a single 2D chemical structure could be written in various forms (for example, ethanol can be denoted as C(O)C, as well as CCO). Thus, we transformed the SMILES strings

extracted from each publication into 2D chemical structures and converted these 2D chemical structures into new SMILES strings using RDKit. Selection of a single group compound identifier or individual compound identifier using the provided drop down menus, affords a two-dimensional rendering of the chemical structure and computed properties of ‘drug-likeness,’ including the number of hydrogen bond donors and acceptors (**Figure 3D**)⁴⁸⁻⁵³. For this section, we created two separate CysDB compound identifiers to produce scatter plots showing the highest ratios collected for each compound.

The final ‘Statistics Section,’ is accessible from the home page both via the chemoproteomics explorer button and from the left menu. The Statistics Section provides interested users with CysDB-wide metrics for hyperreactive and ligandable cysteine-containing proteins, proteins targeted by FDA approved drugs, proteins associated with cancer, and proteins containing missense variants. In a user-centric manner, this section also allows interested users to compare and contrast individual datasets including by identification of unique and overlapping residues and proteins.

(3) Understanding the scope of the CysDB ligandable or putative ‘druggable’ proteome

With the CysDB database established, we further parsed the data available in CysDB to showcase features built into CysDB and to facilitate the identification of new potential targets for future chemical probe development campaigns. More broadly, we also seek to highlight future opportunities for the cysteine chemoproteomic community. Given the aforementioned low overlap between FDA approved drug targets and proteins labeled by cysteine-reactive compounds for prior smaller cysteine chemoproteomics studies¹¹, we next extended this analysis to CysDB. Less than 4% of all human proteins in UniProtKB have been targeted by FDA approved small molecules (**Figure S7**). As only 14.7% of all cysteines in CysDB were reported as likely ligandable, we next performed the same analysis on the subset of proteins in CysDB that contain a ligandable cysteine. Again, consistent with the prior reports that have demonstrated a low overlap between

targets of covalent compounds and FDA approved drugs, we find that 3% of proteins that contain one or more ligandable cysteine have been targeted by FDA approved drugs (**Figure 4A**). Broadening this analysis to a less restrictive set of compound-protein interactions, we find that 32.5% of proteins with ligandable cysteines have been targeted by small-molecules, as reported by ChEMBL, DrugBank, and the FDA (**Figure 4B**). These findings showcase the opportunities for targeting undrugged proteins using cysteine-reactive chemical probes.

Prior studies have shown that drug and putative drug targets are highly enriched for protein classes featuring well defined binding sites, including enzymes and receptors. Therefore, our next step to further characterize whether the CysDB members represent new druggable space was to parse the UniProtKB keyword functional annotations of all ligandable proteins in CysDB. Stratification of the CysDB ligandable proteins into two categories, targeted and untargeted by FDA approved compounds, acknowledged an enrichment for enzymes in the FDA approved subset (**Figure 4C**). In contrast, the functions of the non-FDA subset of ligandable proteins in CysDB span a number of important protein classes, including transcription factors (TFs), which are often categorized as a largely ‘undruggable’ class of proteins, with the notable exception of TFs with well-defined small molecule ligand binding pockets, such as nuclear hormone receptors.

To further dissect the potential druggability of CysDB entries, we next analyzed the compounds that target ligandable cysteine residues. A number of different electrophilic moieties, often termed ‘warheads,’ have been developed, which react with cysteine residues in both irreversible and covalent reversible modes of labeling^{39,54-56}. Examples of these electrophilic handles include compounds that react via a thiol-michael addition (e.g., irreversible modifiers such as acrylamide, fumarate esters, vinyl sulfonamide together with reversible modifiers such as cyanoacrylamide), compounds that react via S_N2 (e.g., α -halo compounds), as well as compounds that react via S_NAr (e.g., halogen-substituted electron deficient heterocycles such as chlorotriazine). As prior studies have revealed varying proteome-wide reactivity and structure-activity relationships (SAR) for different cysteine-reactive electrophiles, we next quantified the

number of cysteines detected as labeled by individual electrophile chemotypes. For this analysis, we defined that a cysteine was labeled by one of the five warheads if the cysteine had an $R \geq 4$ for at least one compound (**Figure 4D**, **Figure S8**, and **Figure S9**)^{2,27-62}. We find that a large majority of the ligandability data were acquired for samples subjected to labeling by acrylamides (AA) and chloroacetamide (CA)-substituted compounds across the panel of cell lines tested (**Figure 4D** and **S10**), with a small fraction derived from additional probes ranging from cyanoacrylamides to dimethylfumarate listed in **Table S2**. Interestingly, we find that some cysteines react promiscuously with both AA and CA electrophiles, whereas others show an electrophile preference (**Figure 4E**). The proteins glutathione S-transferase omega-1 (GSTO1) and carbonyl reductase (CBR1) exemplify the striking electrophile preference observed for some proteins (**Figure 4F**). For GSTO1, the highly ligandable cysteine (Cys 32) exhibits strong preference for reacting with chloroacetamide (CA)-substituted compounds (1 to 11.5 in favor of CA electrophiles, with respect to unique SMILES strings with the CA moiety). In contrast, cysteine 226 of CBR1 shows marked acrylamide (AA) bias (5 to 1 in preference of AA warheads, with respect to unique SMILES strings with the AA moiety).

(4) Characterizing CysDB proteins based on structural, activity and functional annotations

Given the sheer scope of available chemoproteomics datasets, one of the foremost ongoing challenges of cysteine chemoproteomic studies is the high throughput delineation of the functional impact of covalent cysteine modification. While for some cysteines, such as catalytic nucleophiles, covalent modification will almost invariably afford a defined functional outcome, the impact of modifying other less well annotated cysteines, such as those in proteins or protein domains of unknown function, remains less clear. To encourage discovery of likely functional and disease-relevant cysteines, CysDB includes metrics of functionality from UniProtKB, known Cancer Gene Census (CGC), and genetic variants in ClinVar. These databases were chosen to provide measures of relevance to functional biology and human disease.

We first harnessed UniProtKB annotations to determine which CysDB proteins had functional annotations of the following active sites, binding sites, catalytic activity, disulfide bonds and redox potentials. Analysis concluded 1,505 CysDB proteins possess an active site, 2,961 possess a binding site, 2,784 have experimental evidence for catalytic activity, 1,077 have annotated disulfide bonds and 52 have experimental evidence for redox potentials (**Figure 5A**). Comparable distribution of functional annotations was observed when stratifying the CysDB dataset to consider hyperreactive and ligandable proteins.

To assess whether any CysDB cysteines were annotated as known active or binding sites, we parsed the UniProtKB site annotations for residue positions. This analysis uncovered that, while cysteine is a relatively rare amino acid (2.3% of all proteinacious amino acids are cysteines¹), cysteine is the second most abundant binding site amino acid and the third most abundant active site amino acid (**Figure S11** and **Figure S12**). Overall, CysDB reports identification of 1,335 (31.8%) of all known cysteine matching UniProtKB annotated binding sites and 288 (49%) of all known cysteine active sites (**Figure 5B**). Out of the 4,198 cysteine specific binding sites, 178 of them have been liganded by a compound in CysDB. In addition, 98 out of the 583 cysteine active sites have been liganded by a compound in CysDB and 41 out of the 583 cysteine active sites were deemed hyperreactive (**Figure S13**).

Next, we extended this analysis to look for cysteines ‘in or near’ annotated active or binding sites using protein sequences. By searching 10 amino acids upstream and downstream from a CysDB identified cysteine, we were able to increase the number of cysteines proximal to these functional sites. In total, 2,602 CysDB cysteines are near binding sites, including 396 ligandable and 41 hyperreactive CysDB cysteines (**Figure S14**), and 496 CysDB cysteines are near active sites, including 56 ligandable and 12 hyperreactive cysteines (**Figure S15**).

As the UniProtKB dataset is limited to 1D analysis, we next asked whether CysDB could also provide insight into the 3D microenvironment of identified cysteines, using structures reported in the PDB. 5,270 CysDB ID proteins are associated with an available PDB structure, which

represents 70% of all human genes with available crystallographic structures (**Figure S16**). Of these, 2,314 (31%) contain one or more ligandable cysteines and 279 feature at least one hyperreactive cysteine (**Figure 5C**). To confirm whether a CysDB cysteine was resolved in a PDB structure, we parsed the residue numbers and coordinates from PDB files. To account for discrepancies between UniProtKB and PDB residue numbers, residue to protein sequence numbering was mapped using SIFT annotations⁶³ (**Figure S16**). This systematic analysis of residue-level mapping established that out of all the proteins with annotated binding or active sites, 2,684 and 1,315 proteins, respectively, are associated with PDB structures (**Figure S17** and **Figure S18**). Of these, 1,007 proteins have a cysteine binding site resolved in a corresponding structure, while 338 proteins have a cysteine active site resolved in a corresponding structure. In aggregate, 18,959 (30.1%) of CysDB identified cysteines are resolved in a corresponding crystal structure. Further inspection of this dataset revealed that 1,212 CysDB cysteines are proximal (within 10 Angstroms) to binding site residues and 704 CysDB cysteines are proximal to active site residues in 3D space (**Figure S19** and **Figure S20**). To assist structure-guided analysis of cysteine datasets, CysDB provides users with 3D interactive renderings of cysteine-containing structures that include known functional annotations.

Notably, 8,214 proteins (71%) identified by chemoproteomics do not have highly supported evidence in UniProtKB for binding or active sites. Therefore, we next asked whether the CysDB platform could provide additional information about these proteins and corresponding identified cysteines to further aid in delineation of functionally significant cysteines. To guide our platform development efforts, we tested whether the ligandable and hyperreactive cysteine-containing protein subsets are enriched for particular structural domains and functional pathways. Enrichment analysis of protein family (Pfam)⁶⁴ domains elucidated a 13-fold enrichment of liganded proteins in the DEAD/DEAH box helicase family, which is consistent with our prior observation of enrichment for RNA binding proteins in chemoproteomics datasets (**Figure 5D**)⁶⁵.

Responsible for unwinding the duplex of double-stranded RNA, mutations in DEAD/DEAH proteins have been linked to autoimmune disease and some cancers, such as DEAD-Box Helicase 3 X-Linked (DDX3X) in medulloblastoma⁶⁶⁻⁶⁹. Pfam domain enrichment analysis for the hyperreactive cysteine subset, revealed an enrichment of thioredoxin and arginine kinase families. These findings are consistent with prior reports of redox enzymes featuring highly reactive cysteines⁷. Notably creatine kinase enzymes are members of the arginine kinase family of enzymes, which are known to have highly reactive active site cysteines⁷.

We then extended these studies to Panther⁷⁰ pathway analysis to assess if particular pathways are enriched for reactive or ligandable cysteines. We observe an enrichment of ligandable cysteine-containing proteins implicated in apoptosis (**Figure 5E**). Examples of ligandable cysteine-containing proteins include TP53, caspase-8, and APBB2. Given the central relevance in modulating cell death to treatment of numerous disorders, including cancers and neurodegenerative disorders, we expect that this observed notable enrichment indicates untapped opportunities for the development of probes targeting cell death^{71,72}. The hyperreactive cysteine-containing protein set, by contrast, was distinctly enriched for proteins involved in integrin signaling. These findings are consistent with the aforementioned enrichment for hyperreactive cysteines in the thioredoxin proteins and related antioxidant systems that are critical for regulation of integrin abundance, secretion, and disulfide formation^{73,74}.

(5) Stratifying CysDB proteins based on disease-relevant annotations, including cancer association and measures of genetic variation

Building upon our analyses of protein function, we assessed the human disease relevance of the CysDB proteins. Restricting our analysis to the ligandable and hyperreactive subsets, we analyzed which phenotypes were associated with CysDB proteins. Using disease annotations from the Online Mendelian Inheritance in Man (OMIM)⁷⁵ knowledge base, ligandable cysteine-containing proteins showed terms related to a broad range of cancers, including colorectal, breast

and leukemia. The hyperreactive cysteine-containing protein subset was enriched for terms associated with immune-relevant diseases, specifically those affecting the lymphatic system (**Figure S25**). Next, we determined how many CysDB proteins are annotated as cancer driving genes, as dictated by the Cancer Gene Census (CGC)²⁷. 76% of CGC genes have been identified by CysDB (559/733) (**Figure S28**). Out of all the CGC genes, 38% are annotated as ligandable in CysDB, indicating untapped opportunities for the development of tailored therapies targeting driver mutations (**Figure 6A** and **Table S4**). These results compare favorably to the 11% of cancer driving genes that have been targeted by FDA approved small molecules (**Figure S29** and **Table S2**). We observed a considerable difference in the number of available therapies for different cancers during our enrichment analysis for CysDB proteins associated with different tumor types. While acute myeloid leukemia (AML) genes are the most represented somatic tumor type in CGC, only 5% of these genes are targets of FDA approved small molecules. By contrast, 13 out of 38 (34%) of non-small cell lung cancer (NSLC) genes have been targeted by FDA approved drugs. Towards addressing this therapy gap, CysDB detects most CGC genes associated with AML, 71 out of 81 (88%) (**Figure 6B**). In fact, 36 of these AML genes have been liganded by a compound in CysDB, such as class 2 AML genes nucleophosmin 1 (NPM1) and core-binding factor subunit beta (CBFB).

Genetic variants, along with wild-type genes, can contribute towards harmful disease phenotypes. The ClinVar²⁸ database provides a curated set of clinical significance for over a million genetic variants, which are classified as either benign, pathogenic, or variants of unknown significance (VUS). Out of 12,858 unique UniProtKB proteins associated with ClinVar variants (mapped to 31,685 unique genes), 9,478 (73.7%) proteins have a missense variant (**Figure S30**). Overall, more than half of the proteins identified in CysDB have an associated ClinVar missense variant, of which 3,075 contain a liganded cysteine and 330 contain a hyperreactive cysteine (**Figure 6C**). Previously we reported a trend between chemoproteomic identified cysteines and missense pathogenicity, where chemoproteomic detected cysteine codons were predicted to be

more deleterious than undetected cysteine codons⁹. Consistent with the ubiquity of missense variants in ClinVar, the most common mutation associated with CysDB ID CGC genes are missense mutations²⁷. Of the CysDB ID proteins that have a ClinVar missense variant, 4,418 proteins have a benign variant, 2,524 proteins have a pathogenic variant, and 3,333 proteins have a variant of unknown significance (**Figure S31**). The proteins with the highest number of pathogenic variants are Fibrillin-1 (FBN1, UniProtKB: P35555) and Low-density lipoprotein receptor (LDLR, UniProtKB: P01130) (**Figure 6D**). Mutations in FBN1 are known to frequently cause Marfan syndrome by destabilizing disulfide bonds of conserved cysteine residues in epidermal growth factor (EGF)-like domains⁷⁶⁻⁷⁸. Additionally, LDLR contains cysteine-rich repeats that bind lipoproteins. Loss-of-function mutations in these regions result in the disruption of cholesterol transport, leading to an increased risk of heart disease^{78,80}. In addition to enabling human genotype-guided target prioritization, targeting variant-containing chemoproteomic detected proteins may also prove useful precision therapy development in a manner akin to the recent Gly12Cys directed KRAS compounds, including FDA approved Sotorasib⁸¹⁻⁸³.

DISCUSSION

Leading groups in cysteine chemoproteomics have discovered thousands of functional and potentially druggable cysteines proteome-wide¹⁻⁹. These studies have yielded global measures of the SAR of compounds that target specific cysteines together with the intrinsic reactivity towards promiscuous electrophilic probes. Given the functional and clinical significance of identification of reactive and ligandable cysteines, the development of strategies that enable rapid cross datasets comparisons between these studies represents an important opportunity for the cysteine chemoproteomics community that will enable a more comprehensive understanding of the cysteinome. Here we present CysDB as such a tool that unites high coverage chemoproteomic measures of identification, ligandability, and hyperreactivity across multiple studies, together with integration with relevant resources to provide metrics of functionality and

disease-relevance. CysDB achieves identification of an impressive 62,888 unique cysteines and 11,621 proteins, which represents a ~100% increase in total number of identified cysteine residues compared to individual prior studies, with added potential for further growth as new datasets become available.

As a first step to construct CysDB, we accumulated and curated a selected set of cysteine chemoproteomics studies, which were prioritized due to the high coverage of identified cysteines. During our stringent data curation, we observed study-dependent differences in conventions for designating a cysteine as hyperreactive and/or ligandable. To account for the potential uncertainty caused by a general absence of field-wide data analysis conventions, we retained all hyperreactive and/or liganded cysteines so as to accurately represent each study's reported findings. The development of statistically rigorous conventions for the field will aid in normalizing future cross-dataset comparison efforts. As a first approach, in our studies we have required comparable ratios with low standard deviations identified across multiple biological replicates together with inclusion of inactive control datasets to further simplify removal of potentially spurious elevated ratios. For studies that rely on MS1-based quantification, so-called 'singleton' values, should be treated with an additional level of stringency, as these can prove more prone to yielding spurious ratios. These ratios are derived from peptides with precursor ions that have only been identified with either a heavy or light isotopic modification. Therefore, we followed general conventions for filtering singletons, by setting a maximum ratio value of $\log_2(\text{ratio})$ equivalent to 20 requiring identification of additional lower ratio ions. Future studies, including our own, will benefit significantly from harnessing advances in data acquisition and analysis to improve reproducibility, including imputation and data independent acquisition (DIA), as showcased by recent efforts by the Wang group⁸⁴.

Illustrating the utility of CysDB, we find that by combining datasets generated across multiple cell lines and using different labeling reagents, we substantially increased aggregate coverage of the cysteinome. Alongside cysteine coverage, CysDB reveals that cell line selection

can impact not only which cysteines are identified in proteomes derived from different cell lines (**Figure S5**), but also the hyperreactivity and ligandability of individual cysteines. We ascribe these differences in part to both cell state specific expression as well as the stochastic nature of data dependent acquisition (DDA), which is the acquisition method used to generate nearly all datasets analyzed.

In its current iteration, CysDB provides a low-throughput mechanism to assess reproducible ligandability of cysteines across studies, including those that analyze identical compounds. To enable such comparisons, we grouped identical compounds shared across multiple publication datasets under a shared identifier, termed "Group Compound ID." The Group Compound ID allows users to easily visualize the reproducibility of cysteine ligandability across studies. The relative rarity of shared compounds used across multiple studies (25 in total in CysDB) remains a limitation for reproducibility analysis at the level of specific compounds. One notable exception to this paradigm is the recent work by Yang et al.¹⁰ that validates many compounds assayed by DDA using a DIA approach. We hope that future studies will consider inclusion of several benchmark scout fragments to stimulate efforts in assessing the reproducibility of ligandable ratios across studies. In addition, these cross-dataset comparisons revealed a marked bias towards chemoproteomic analysis of chloroacetamide and acrylamides, which points to largely untapped opportunities in expanding the scope of the ligandable cysteinome through assaying additional classes of electrophiles.

A key feature of CysDB is the inclusion of functional and disease annotations from UniProtKB, CGC, and ClinVar. We expect that the centralization of the annotations should allow for rapid prioritization of ligandable cysteines for future studies. Showcasing the utility of cysteine chemoproteomics to access tough-to-drug classes of proteins, we find a considerable enrichment in transcription factors containing ligandable cysteines (**Figure 4C**). We also observe that the vast majority of Census driver genes contain a cysteine identified in a chemoproteomics study. These findings together with our observation that a smaller but still substantial 38% of all census genes

contain a ligandable cysteine suggests opportunities for future studies to more comprehensively assess the ligandability of these genes.

During our efforts to map annotations generated from genomics data (e.g., ClinVar/Census data), we encountered issues with mismapping for a subset of identifiers. While processing all datasets included in CysDB, we observed that a handful (16) gene names did not map to UniProtKB protein accession numbers in a one-to-one type of manner, during SQL querying; multiple HGNC or Gene Entrez symbols can be associated with a single protein identifier if the translated gene products are identical protein sequences²⁶. Given the utility of a gene-centric search, we have incorporated such identifiers in this release of CysDB to aid future proteogenomic analysis.

An ongoing goal of CysDB is to facilitate expanding the scope of the ligandable and potentially druggable cysteinome, particularly for functional and disease-relevant proteins. Given our observed bias in CysDB ligandability datasets towards chloroacetamide and acrylamide moieties, we expect that future expansions of the ligandable cysteinome may stem in part from chemoproteomic studies utilizing additional classes of electrophiles. In a similar manner, we expect that inclusion of datasets generated using alternatives to iodoacetamide as promiscuous cysteine-reactive capping agents, including for example hypervalent iodine-based probes¹⁹, should further increase coverage of labeled cysteines. In this first iteration of CysDB, we have opted to restrict our datasets to those generated through lysate-based proteomic studies, which eliminates challenges associated with deconvolving changes in protein abundance from direct cysteine labeling. Given the importance of cell-based studies for target discovery and hit-to-lead optimization, we look forward to including such datasets in future releases, particularly when combined with bulk measures of protein abundance. In a similar manner, we look forward to incorporating redox proteomics datasets in subsequent iterations of CysDB, alongside generalized strategies to merge the diverse data formats generated by these studies. Looking ahead, we are enthusiastic about the continued growth of CysDB and encourage all interested

users to consider submission of relevant chemoproteomics datasets that comply with our submission format (**Table S1**) and that include spectral files deposited in a public data repository, such as Pride⁸⁵.

STAR METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- METHOD DETAILS
 - Proteomics data analysis
 - CysDB database
 - CysDB web application
- DATA AND CODE AVAILABILITY
- ADDITIONAL RESOURCES
 - Dataset Addition to CysDB

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
CysDB	This paper	https://backuslab.shinyapps.io/cysdb/
Human proteome	UniProt	UP000005640
UniProtKB/Swiss-Prot Fasta	2201-release	https://www.uniprot.org/
UniProtKB/Swiss-Prot	2209-release	https://www.uniprot.org/
COSMIC	2209-release	https://cancer.sanger.ac.uk/census

ClinVar	2209-release	https://cancer.sanger.ac.uk/census
Human Protein Atlas (HPA)	Version 21.1	https://www.proteinatlas.org
Enrichr Panther	2016	http://www.pantherdb.org/pathway/
Enrichr Pfam Domains	2019	https://pfam.xfam.org/
Enrichr OMIM Disease		https://www.omim.org/downloads
Software and Algorithms		
R version 4.2.1	R project	https://www.r-project.org
RStudio Version 2022.07.1		https://rstudio.com
ImageJ	NIH	https://imagej.nih.gov/ij/
Enrichr	Accessed Sept. 2022	https://maayanlab.cloud/Enrichr/

LEAD CONTACT AND MATERIALS AVAILABILITY

Materials and data in this work can be obtained from Keriann Backus (kbackus@mednet.ucla.edu) upon request.

METHODS DETAILS

Proteomics data analysis

Chemoproteomics data was collected from publicly accessible supplementary tables of previous literature^{2,4-11}. Columns were parsed for UniProtKB protein identifiers and locations of the corresponding modified cysteine amino acid numbers to create a new identifier for CysDB:

UniProtKBID_CYS#. Any cysteine classified as 'ligandable' or 'hyperreactive' is listed in CysDB as ligandable or hyperreactive. Individual ligandability and reactivity ratios found from each publication are listed in **Tables S1** and **Table S2**. In some cases, for the ligandability and reactivity datasets, publications listed ratios for peptides simultaneously modified at multiple cysteines such as UniProtKBID_CYS#1_CYS#2, where the ratios provided for UniProtKBID_CYS#1_CYS#2 differed from UniProtKBID_CYS#1. Thus, ratios for peptides modified at multiple cysteines were not included in further analyses.

Compounds found in ligandability studies were stratified according to their cell line and chemotype. Unique identifiers for each compound were constructed based on their chemotype within the five categories: acrylamide, bromoacetamide, chloroacetamide, dimethyl fumarate (dmf) and others, such as ACRYL_#. Unique group identification numbers were constructed for compounds based on their chemotype and SMILES string, such as GROUP_ACRYL_# Publication names for each compound and CysDB names are provided in **Table S2**.

In the event amino acid numbers were not provided by the author, python scripts (available on GitHub) were utilized to map the listed peptide sequences to the canonical protein sequences of the 2201-release UniProtKB human fasta reference file, as this release is the only version saved in the UniProtKB archive for future mapping. Cysteines from unmatched peptides were removed prior to subsequent analyses. To inspect the extent of mismapped identifiers in CysDB, we collected peptides mapped to multiple proteins or peptides labeled at multiple cysteine sites from each publication (**Table S1**). Peptides labeled at multiple cysteines were dropped from our ligandability and hyperreactivity data aggregation.

Cancer Gene Census (CGC) website reports were downloaded Sept. 2022 and mapped to CysDB data using UniProtKB accessions. Due to frequent UniProtKB updates, Gene symbols reported

in the Cancer Gene Census were mapped to gene names in UniProtKB to identify the updated UniProtKB codes (2209-release).

CysDB database

CysDB was created as a relational database using MySQL v.8.0. Overall, the database contains six tables and is hosted on Google Cloud. The major parent tables, 'Datasets' and 'Identifiers', were further broken down into child tables, such as 'Ligandable', 'Reactive', 'Compound' and 'Warheads' (**Figure S6**). The Datasets table contains information specific to each of the nine publications, while the Identifiers table contains information specific to each modified cysteine or protein identifier. Columns within Datasets and Identifiers include binary results for the following three categories: identified, hyperreactive and ligandable. However, individual competition ratios are listed in the Ligandable table and individual reactivity ratios are listed in the Reactive table. Calculated molecular properties for 'drug-likeness' were acquired using RDKit45 and are stored in the 'Compounds' table. This table also contains the CysDB compound identifier mapped to their associated publication abbreviation or designated name. Group compound identifiers ("GROUP_WARHEAD_#") were defined by unique standardized SMILES strings and individual compound identifiers ("WARHEAD_#") were defined by unique standardized SMILES string, cell line and publication author combinations. Finally, the warhead table holds chemotype classifications for each compound. The five chemotype classifications were as follows: acrylamide, bromoacetamide, chloroacetamide, dimethyl fumarate and other.

CysDB web application

The CysDB web application was developed using the Shiny R package (<https://shiny.rstudio.com/>). Schematics of protein sequence chains, domains and motifs on the CysDB web server are constructed using the drawProteins R package (<https://github.com/brennanpincardiff/drawProteins>). Interactive viewing of PDB crystal structures

is performed using NGLViewR (<https://github.com/nglviewer/nglview>). Protein protein interaction networks are accessed via the STRING database (<https://string-db.org/>). Gene set library enrichment analyses are provided with the Enrichr R package (<https://maayanlab.cloud/Enrichr/>) and ontology enrichment plots are produced with the gprofiler2 R package (<https://biit.cs.ut.ee/gprofiler/gost>). All plots are generated with the ggplot2 and plotly (<https://plotly.com/r/>) R libraries.

DATA AND CODE AVAILABILITY

The dataset and source code are available at https://github.com/lmboat/cysdb_app.

ADDITIONAL RESOURCES

The CysDB dataset is provided as an interactive web resource at <https://backuslab.shinyapps.io/cysdb/>.

Dataset Addition to CysDB Guidelines

Email submission materials to cysteineomedb@gmail.com with the following information: copy of publication, supplementary information, additional details for data filtering and note the version of UniProt used to obtain protein accessions. Proteins must be identified through UniProtKB accessions. Please use the format, UniProtKBID_CYS#, to indicate which residues have been labeled. For ligandability experiments using a variety of electrophiles, inclusion of SMILES strings and criteria for 'ligandability' classification is required (ex. $R \geq 4$ for at least n number of compounds). Table templates and additional information for submission requests can be found in **Table S1**.

Significance

Chemoproteomics has emerged as highly enabling technology capable of pinpointing functional and potentially druggable cysteine residues proteome-wide. While many cysteine chemoproteomic datasets are now available, the overlap and reproducibility between studies remains unknown due to a lack of mechanisms for data integration. Here we report CysDB, a comprehensive database of cysteine chemoproteomics data that facilitates rapid discovery of the reactivity, ligandability, potential therapeutic relevance for 62,888 cysteines and 11,621 proteins. By including available data from multiple published studies together with annotations of function, disease relevance, and structural information, CysDB represents an unparalleled resource for understanding the scope and functional significance of the cysteinome. Given the emerging value of cysteine-reactive molecules as clinical candidates and chemical probes, CysDB also provides a resource for ongoing and future electrophilic probe development campaigns.

Acknowledgements

This study was supported by a Beckman Young Investigator Award (K.M.B.), DOD-Advanced Research Projects Agency (DARPA) D19AP00041 (K.M.B.), and NIGMS System and Integrative Biology 5T32GM008185-33 (L.M.B.). We thank all members of the Backus lab for helpful suggestions. We thank S. Forli and J. Eberhardt for helpful suggestions.

Author Contributions

L.M.B., D.K.S. and K.M.B. conceived of the project. L.M.B. and M.F.P performed data analysis. L.M.B wrote software and created the database. D.K.S. provided technical advice. L.M.B. and K.M.B. wrote the manuscript.

Conflicts of Interest

The authors declare no financial or commercial conflict of interest.

REFERENCES

- (1) Xiao, H.; Jedrychowski, M. P.; Schweppe, D. K.; Huttlin, E. L.; Yu, Q.; Heppner, D. E.; Chouchani, E. T. A Quantitative Tissue-Specific Landscape of Protein Redox Regulation during Aging. *Cell* **2020**, *180* (5), 968–983.
- (2) 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* (5), 630–641.
- (3) Müller, S.; Ackloo, S.; Al Chawaf, A.; Al-Lazikani, B.; Antolin, A.; Baell, J. B.; Arrowsmith, C. H. Target 2035—Update on the Quest for a Probe for Every Protein. *RSC Med. Chem.* **2022**, *13* (1), 13–21.
- (4) Yan, T.; Desai, H. S.; Boatner, L. M.; Yen, S. L.; Cao, J.; Palafox, M. F.; Jami-Alahmadi, Y.; Backus, K. SP3-FAIMS Chemoproteomics for High Coverage Profiling of the Human Cysteinome. *ChemBioChem* **2021**.
- (5) 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* (4), 2610–2618. <https://doi.org/10.1021/acs.analchem.0c04726>.
- (6) Li, Z.; Liu, K.; Xu, P.; Yang, J. Benchmarking Cleavable Biotin Tags for Peptide-Centric Chemoproteomics. *J. Proteome Res.* **2022**, *21* (5), 1349–1358. <https://doi.org/10.1021/acs.jproteome.2c00174>.
- (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* (7325), 790–797. <https://doi.org/10.1038/nature09472>.
- (8) 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.; Cravatt, B. F. An Activity-Guided Map of Electrophile-Cysteine Interactions in Primary Human T Cells. *Cell* **2020**, *182* (4), 1009–1026.
- (9) Palafox, M. F.; Desai, H. S.; Arboleda, V. A.; Backus, K. M. From Chemoproteomic-detected Amino Acids to Genomic Coordinates: Insights into Precise Multi-omic Data Integration. *Mol. Syst. Biol.* **2021**, *17* (2). <https://doi.org/10.1525/msb.20209840>.
- (10) 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* (2), 901–911. <https://doi.org/10.1021/jacs.1c11053>.
- (11) 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* (7608), 570–574.
- (12) Bar-Peled, L.; Kemper, E. K.; Suciu, R. M.; Vinogradova, E. V.; Backus, K. M.; Horning, B. D.; Cravatt, B. F. Chemical proteomics identifies druggable vulnerabilities in a genetically defined cancer. *Cell* **2017**, *171* (3), 696–709.

- (13) Backus, K. M. *Applications of Reactive Cysteine Profiling*; Activity-Based Protein Profiling, 2018.
- (14) 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 Benzenedioxolone Reagents. *Angew. Chem.* **2015**, *127* (37), 11002–11007.
- (15) Kulkarni, R. A.; Bak, D. W.; Wei, D.; Bergholtz, S. E.; Briney, C. A.; Shrimp, J. H.; Meier, J. L. A chemoproteomic portrait of the oncometabolite fumarate. *Nat. Chem. Biol.* **2019**, *15* (4), 391–400.
- (16) Grossman, E. A.; Ward, C. C.; Spradlin, J. N.; Bateman, L. A.; Huffman, T. R.; Miyamoto, D. K.; Nomura, D. K. Covalent Ligand Discovery against Druggable Hotspots Targeted by Anti-Cancer Natural Products. *Cell Chem. Biol.* **2017**, *24* (11), 1368–1376.
- (17) Tian, C.; Sun, R.; Liu, K.; Fu, L.; Liu, X.; Zhou, W.; Yang, J. Multiplexed Thiol Reactivity Profiling for Target Discovery of Electrophilic Natural Products. *Cell Chem. Biol.* **2017**, *24* (11), 1416–1427.
- (18) 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* (1), 79–85.
- (19) Abegg, D.; Tomanik, M.; Qiu, N.; Pechalrieu, D.; Shuster, A.; Commare, B.; Adibekian, A. Chemoproteomic Profiling by Cysteine Fluoroalkylation Reveals Myrocin G as an Inhibitor of the Nonhomologous End Joining DNA Repair Pathway. *J. Am. Chem. Soc.* **2021**, *143* (48), 20332–20342.
- (20) Fu, L.; Li, Z.; Liu, K.; Tian, C.; He, J.; He, J.; He, F.; Xu, P.; Yang, J. A Quantitative Thiol Reactivity Profiling Platform to Analyze Redox and Electrophile Reactive Cysteine Proteomes. *Nat. Protoc.* **2020**, *15* (9), 2891–2919. <https://doi.org/10.1038/s41596-020-0352-2>.
- (21) Desai, H. S.; Yan, T.; Yu, F.; Sun, A. W.; Villanueva, M.; Nesvizhskii, A. I.; Backus, K. M. SP3-Enabled Rapid and High Coverage Chemoproteomic Identification of Cell-State-Dependent Redox-Sensitive Cysteines. *Mol. Cell. Proteomics* **2022**, *21* (4), 100218.
- (22) Shi, Y.; Fu, L.; Yang, J.; Carroll, K. S. Wittig Reagents for Chemoselective Sulfenic Acid Ligation Enables Global Site Stoichiometry Analysis and Redox-Controlled Mitochondrial Targeting. *Nat. Chem.* **2021**, *13* (11), 1140–1150.
- (23) Mnatsakanyan, R.; Markoutsas, S.; Walbrunn, K.; Roos, A.; Verhelst, S. H.; Zahedi, R. P. Proteome-Wide Detection of S-Nitrosylation Targets and Motifs Using Bioorthogonal Cleavable-Linker-Based Enrichment and Switch Technique. *Nat. Commun.* **2019**, *10* (1), 1–12.
- (24) Wu, S.; Luo, H.; Wang, H.; Zhao, W.; Hu, Q.; Yang, Y. Cysteinome: The First Comprehensive Database for Proteins with Targetable Cysteine and Their Covalent Inhibitors. *Biochem. Biophys. Res. Commun.* **2016**, *478* (3), 1268–1273.
- (25) Yan, T.; Palmer, A. B.; Geiszler, D. J.; Polasky, D. A.; Boatner, L. M.; Burton, N. R.; Armenta, E.; Nesvizhskii, A. I.; Backus, K. M. Enhancing Cysteine Chemoproteomic Coverage through Systematic Assessment of Click Chemistry Product Fragmentation. *Anal. Chem.* **2022**, *94* (9), 3800–3810. <https://doi.org/10.1021/acs.analchem.1c04402>.
- (26) Consortium, U. UniProt: A Worldwide Hub of Protein Knowledge. *Nucleic Acids Res.* **2019**, *47* (D1), 506–515.
- (27) Sondka, Z.; Bamford, S.; Cole, C. G.; Ward, S. A.; Dunham, I.; Forbes, S. A. The COSMIC Cancer Gene Census: Describing Genetic Dysfunction across All Human Cancers. *Nat. Rev. Cancer* **2018**, *18* (11), 696–705.

- (28) Landrum, M. J. ClinVar: Improving Access to Variant Interpretations and Supporting Evidence. *Nucleic Acids Res.* **2018**, *46* (D1), 1062–1067.
- (29) Uhlen, M. Towards a Knowledge-Based Human Protein Atlas. *Nat. Biotechnol.* **2010**, *28* (12), 1248–1250.
- (30) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; Veij, M.; Félix, E.; Leach, A. R. ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* **2019**, *47* (D1), 930–940.
- (31) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Wilson, M. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* **2018**, *46* (D1), 1074–1082.
- (32) Rose, P. W.; Prlić, A.; Altunkaya, A.; Bi, C.; Bradley, A. R.; Christie, C. H.; Burley, S. K. The RCSB Protein Data Bank: Integrative View of Protein, Gene and 3D Structural Information. *Nucleic Acids Res.* **2016**, *gkw1000*.
- (33) Eng, J. K.; McCormack, A. L.; Yates, J. R. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (11), 976–989. [https://doi.org/10.1016/1044-0305\(94\)80016-2](https://doi.org/10.1016/1044-0305(94)80016-2).
- (34) Yu, F.; Teo, G. C.; Kong, A. T.; Haynes, S. E.; Avtonomov, D. M.; Geiszler, D. J.; Nesvizhskii, A. I. Identification of Modified Peptides Using Localization-Aware Open Search. *Nat. Commun.* **2020**, *11* (1), 4065. <https://doi.org/10.1038/s41467-020-17921-y>.
- (35) Integrated Proteomics Pipeline (IP2). <http://goldfish.scripps.edu/>.
- (36) Xu, T.; Park, S. K.; Venable, J. D.; Wohlschlegel, J. A.; Diedrich, J. K.; Cociorva, D.; Lu, B.; Liao, L.; Hewel, J.; Han, X.; Wong, C. C. L.; Fonslow, B.; Delahunty, C.; Gao, Y.; Shah, H.; Yates, J. R. ProLuCID: An Improved SEQUEST-like Algorithm with Enhanced Sensitivity and Specificity. *J. Proteomics* **2015**, *129*, 16–24. <https://doi.org/10.1016/j.jprot.2015.07.001>.
- (37) Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I. MSFragger: Ultrafast and Comprehensive Peptide Identification in Mass Spectrometry–Based Proteomics. *Nat. Methods* **2017**, *14* (5), 513–520. <https://doi.org/10.1038/nmeth.4256>.
- (38) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: An Open-Source MS/MS Sequence Database Search Tool. *PROTEOMICS* **2013**, *13* (1), 22–24. <https://doi.org/10.1002/pmic.201200439>.
- (39) Serafimova, I. M.; Pufall, M. A.; Krishnan, S.; Duda, K.; Cohen, M. S.; Maglathlin, R. L.; Taunton, J. Reversible Targeting of Noncatalytic Cysteines with Chemically Tuned Electrophiles. *Nat. Chem. Biol.* **2012**, *8* (5), 471–476.
- (40) Hacker, S. M.; Backus, K. M.; Lazear, M. R.; Forli, S.; Correia, B. E.; Cravatt, B. F. Global Profiling of Lysine Reactivity and Ligandability in the Human Proteome. *Nat. Chem.* **2017**, *9* (12), 1181–1190.
- (41) Abbasov, M. E.; Kavanagh, M. E.; Ichu, T. A.; Lazear, M. R.; Tao, Y.; Crowley, V. M.; Cravatt, B. F. A Proteome-Wide Atlas of Lysine-Reactive Chemistry. *Nat. Chem.* **2021**, *13* (11), 1081–1092.
- (42) Braschi, B.; Denny, P.; Gray, K.; Jones, T.; Seal, R.; Tweedie, S.; Bruford, E. Genenames. Org: The HGNC and VGNC Resources in 2019. *Nucleic Acids Res.* **2019**, *47* (D1), 786–792.
- (43) Consortium, G. O. The Gene Ontology Resource: 20 Years and Still GOing Strong. *Nucleic Acids Res.* **2019**, *47* (D1), 330–338.

- (44) Fabregat, A.; Jupe, S.; Matthews, L.; Sidiropoulos, K.; Gillespie, M.; Garapati, P.; D'Eustachio, P. The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **2018**, *46* (D1), 649–655.
- (45) Chen, E. Y.; Tan, C. M.; Kou, Y.; Duan, Q.; Wang, Z.; Meirelles, G. V.; Clark, N. R.; Ma'ayan, A. Enrichr: Interactive and Collaborative HTML5 Gene List Enrichment Analysis Tool. *BMC Bioinformatics* **2013**, *128* (14).
- (46) Kuleshov, M. V.; Jones, M. R.; Rouillard, A. D.; Fernandez, N. F.; Duan, Q.; Wang, Z.; Koplev, S.; Jenkins, S. L.; Jagodnik, K. M.; Lachmann, A.; McDermott, M. G.; Monteiro, C. D.; Gundersen, G. W.; Ma'ayan, A. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **2016**, *gkw377*.
- (47) Schoenmaker, L.; Béquignon, O. J.; Jespers, W.; Westen, G. J. UnCorrupt SMILES: A Novel Approach to de Novo Design, 2022.
- (48) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the Chemical Beauty of Drugs. *Nat. Chem.* **2012**, *4* (2), 90–98.
- (49) Benet, L. Z.; Hosey, C. M.; Ursu, O.; Oprea, T. I. BDDCS, the Rule of 5 and Druggability. *Adv. Drug Deliv. Rev.* **2016**, *101*, 89–98.
- (50) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **2012**, *64*, 4–17.
- (51) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. *J. Comb. Chem.* **1999**, *1* (1), 55–68.
- (52) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A rule of Three for Fragment-Based Lead Discovery? *Drug Discov. Today* **2003**, *8* (19), 876–877.
- (53) Landrum, G. Rdkit Documentation. *Release* **2013**, *1* (1–79), 4.
- (54) Senkane, K.; Vinogradova, E. V.; Suci, R. M.; Crowley, V. M.; Zaro, B. W.; Bradshaw, J. M.; Cravatt, B. F. The Proteome-Wide Potential for Reversible Covalency at Cysteine. *Angew. Chem.* **2019**, *131* (33), 11507–11511.
- (55) Krishnan, S.; Miller, R. M.; Tian, B.; Mullins, R. D.; Jacobson, M. P.; Taunton, J. Design of Reversible, Cysteine-Targeted Michael Acceptors Guided by Kinetic and Computational Analysis. *J. Am. Chem. Soc.* **2014**, *136* (36), 12624–12630.
- (56) Zambaldo, C.; Vinogradova, E. V.; Qi, X.; Iaconelli, J.; Suci, R. M.; Koh, M.; Bollong, M. J. 2-Sulfonylpyridines as Tunable, Cysteine-Reactive Electrophiles. *J. Am. Chem. Soc.* **2020**, *142* (19), 8972–8979.
- (57) Du, X.; Guo, C.; Hansell, E.; Doyle, P. S.; Caffrey, C. R.; Holler, T. P.; Cohen, F. E. Synthesis and Structure–Activity Relationship Study of Potent Trypanocidal Thio Semicarbazone Inhibitors of the Trypanosomal Cysteine Protease Cruzain. *J. Med. Chem.* **2002**, *45* (13), 2695–2707.
- (58) Greenbaum, D. C.; Mackey, Z.; Hansell, E.; Doyle, P.; Gut, J.; Caffrey, C. R.; Chibale, K. Synthesis and Structure–Activity Relationships of Parasiticidal Thiosemicarbazone Cysteine Protease Inhibitors against Plasmodium Falciparum, Trypanosoma Brucei, and Trypanosoma Cruzi. *J. Med. Chem.* **2004**, *47* (12), 3212–3219.

- (59) Shenai, B. R.; Lee, B. J.; Alvarez-Hernandez, A.; Chong, P. Y.; Emal, C. D.; Neitz, R. J.; Rosenthal, P. J. Structure-Activity Relationships for Inhibition of Cysteine Protease Activity and Development of Plasmodium Falciparum by Peptidyl Vinyl Sulfones. *Antimicrob. Agents Chemother.* **2003**, *47* (1), 154–160.
- (60) Klüver, E.; Schulz-Maronde, S.; Scheid, S.; Meyer, B.; Forssmann, W. G.; Adermann, K. Structure–activity Relation of Human β -Defensin 3: Influence of Disulfide Bonds and Cysteine Substitution on Antimicrobial Activity and Cytotoxicity. *Biochemistry* **2005**, *44* (28), 9804–9816.
- (61) Grzonka, Z.; Jankowska, E.; Kasprzykowski, F.; Kasprzykowska, R.; Lankiewicz, L.; Wiczak, W.; Grubb, A. Structural studies of cysteine proteases and their inhibitors. *Acta Biochim. Pol.* **2001**, *48* (1), 1–20.
- (62) Zanon, P. R.; Yu, F.; Musacchio, P.; Lewald, L.; Zollo, M.; Krauskopf, K.; Hacker, S. M. Profiling the Proteome-Wide Selectivity of Diverse Electrophiles, 2021.
- (63) Dana, J. M.; Gutmanas, A.; Tyagi, N.; Qi, G.; O'Donovan, C.; Martin, M.; Velankar, S. SIFTS: Updated Structure Integration with Function, Taxonomy and Sequences Resource Allows 40-Fold Increase in Coverage of Structure-Based Annotations for Proteins. *Nucleic Acids Res.* **2019**, *47* (D1), 482–489.
- (64) Mistry, J.; Chuguransky, S.; Williams, L.; Qureshi, M.; Salazar, G. A.; Sonnhammer, E. L.; Bateman, A. Pfam: The Protein Families Database in 2021. *Nucleic Acids Res.* **2021**, *49* (D1), 412–419.
- (65) Julio, A. R.; Backus, K. M. New Approaches to Target RNA Binding Proteins. *Curr. Opin. Chem. Biol.* **2021**, *62*, 13–23.
- (66) Cruz, J.; Kressler, D.; Linder, P. Unwinding RNA in Saccharomyces Cerevisiae: DEAD-Box Proteins and Related Families. *Trends Biochem. Sci.* **1999**, *24* (5), 192–198.
- (67) Aubourg, S.; Kreis, M.; Lecharny, A. The DEAD Box RNA Helicase Family in Arabidopsis Thaliana. *Nucleic Acids Res.* **1999**, *27* (2), 628–636.
- (68) Patmore, D. M.; Jassim, A.; Nathan, E.; Gilbertson, R. J.; Tahan, D.; Hoffmann, N.; Gilbertson, R. J. DDX3X suppresses the susceptibility of hindbrain lineages to medulloblastoma. *Dev. Cell* **2020**, *54* (4), 455–470.
- (69) Andrisani, O.; Liu, Q.; Kehn, P.; Leitner, W. W.; Moon, K.; Vazquez-Maldonado, N.; Gale, M. Biological Functions of DEAD/DEAH-Box RNA Helicases in Health and Disease, 2022.
- (70) Mi, H.; Muruganujan, A.; Ebert, D.; Huang, X.; Thomas, P. D. PANTHER Version 14: More Genomes, a New PANTHER GO-Slim and Improvements in Enrichment Analysis Tools. *Nucleic Acids Res.* **2019**, *47* (D1), 419–426.
- (71) Fesik, S. W. Promoting Apoptosis as a Strategy for Cancer Drug Discovery. *Nat. Rev. Cancer* **2005**, *5* (11), 876–885.
- (72) Aguilar, A.; Lu, J.; Liu, L.; Du, D.; Bernard, D.; McEachern, D.; Wang, S. Discovery of 4-((3' R, 4' S, 5' R)-6''-Chloro-4'-(3-Chloro-2-Fluorophenyl)-1'-Ethyl-2''-Oxodispiro [Cyclohexane-1, 2'-Pyrrolidine-3', 3''-Indoline]-5'-Carboxamido) Bicyclo [2.2. 2] Octane-1-Carboxylic Acid (AA-115/APG-115): A Potent and Orally Active Murine Double Minute 2 (MDM2) Inhibitor in Clinical Development. *J. Med. Chem.* **2017**, *60* (7), 2819–2839.
- (73) Giancotti, F. G.; Ruoslahti, E. Integrin Signaling. *Science* **1999**, *285* (5430), 1028–1033.
- (74) Cooper, J.; Giancotti, F. G.; A., S.; F., A.; Amberger, J. S.; Bocchini, C. A.; McKusick, V. A. Integrin Signaling in Cancer: Mechanotransduction, Stemness, Epithelial Plasticity, and Therapeutic Resistance. *Cancer Cell* **2019**, *35* (3), 347–367 .,

- (75) Hamosh, A. Online Mendelian Inheritance in Man (OMIM), a Knowledgebase of Human Genes and Genetic Disorders. *Nucleic Acids Res.* **2004**, 33 (Database issue), D514–D517. <https://doi.org/10.1093/nar/gki033>.
- (76) Schrijver, I.; Liu, W.; Brenn, T.; Furthmayr, H.; Francke, U. Cysteine Substitutions in Epidermal Growth Factor-like Domains of Fibrillin-1: Distinct Effects on Biochemical and Clinical Phenotypes. *Am. J. Hum. Genet.* **1999**, 65 (4), 1007–1020.
- (77) Russell, D. W.; Brown, M. S.; Goldstein, J. L. Different Combinations of Cysteine-Rich Repeats Mediate Binding of Low Density Lipoprotein Receptor to Two Different Proteins. *J. Biol. Chem.* **1989**, 264 (36), 21682–21688.
- (78) Daly, N. L.; Scanlon, M. J.; Djordjevic, J. T.; Kroon, P. A.; Smith, R. Three-Dimensional Structure of a Cysteine-Rich Repeat from the Low-Density Lipoprotein Receptor. *Proc. Natl. Acad. Sci.* **1995**, 92 (14), 6334–6338.
- (79) Esser, V.; Limbird, L. E.; Brown, M. S.; Goldstein, J. L.; Russell, D. W. Mutational Analysis of the Ligand Binding Domain of the Low Density Lipoprotein Receptor. *J. Biol. Chem.* **1988**, 263 (26), 13282–13290.
- (80) Lanman, B. A.; Allen, J. R.; Allen, J. G.; Amegadzie, A. K.; Ashton, K. S.; Booker, S. K.; Cee, V. J. Discovery of a Covalent Inhibitor of KRASG12C (AMG 510) for the Treatment of Solid Tumors, 2019.
- (81) Janes, M. R.; Zhang, J.; Li, L. S.; Hansen, R.; Peters, U.; Guo, X.; Liu, Y. Targeting KRAS Mutant Cancers with a Covalent G12C-Specific Inhibitor. *Cell* **2018**, 172 (3), 578–589.
- (82) Patricelli, M. P.; Janes, M. R.; Li, L. S.; Hansen, R.; Peters, U.; Kessler, L. V.; Liu, Y. Selective Inhibition of Oncogenic KRAS Output with Small Molecules Targeting the Inactive State Targeting Inactive KRASG12C Suppresses Oncogenic Signaling. *Cancer Discov.* **2016**, 6 (3), 316–329.
- (83) Ostrem, J. M.; Peters, U.; Sos, M. L.; Wells, J. A.; Shokat, K. M. K-Ras (G12C) Inhibitors Allosterically Control GTP Affinity and Effector Interactions. *Nature* **2013**, 503 (7477), 548–551.
- (84) Tyanova, S.; Temu, T.; Sinitcyn, P.; Carlson, A.; Hein, M. Y.; Geiger, T.; Cox, J. The Perseus Computational Platform for Comprehensive Analysis of (Prote) Omics Data. *Nat. Methods* **2016**, 13 (9), 731–740.
- (85) Perez-Riverol, Y.; Bai, J.; Bandla, C.; García-Seisdedos, D.; Hewapathirana, S.; Kamatchinathan, S.; Vizcaíno, J. A. The PRIDE Database Resources in 2022: A Hub for Mass Spectrometry-Based Proteomics Evidences. *Nucleic Acids Res.* **2022**, 50 (D1), 543–552.

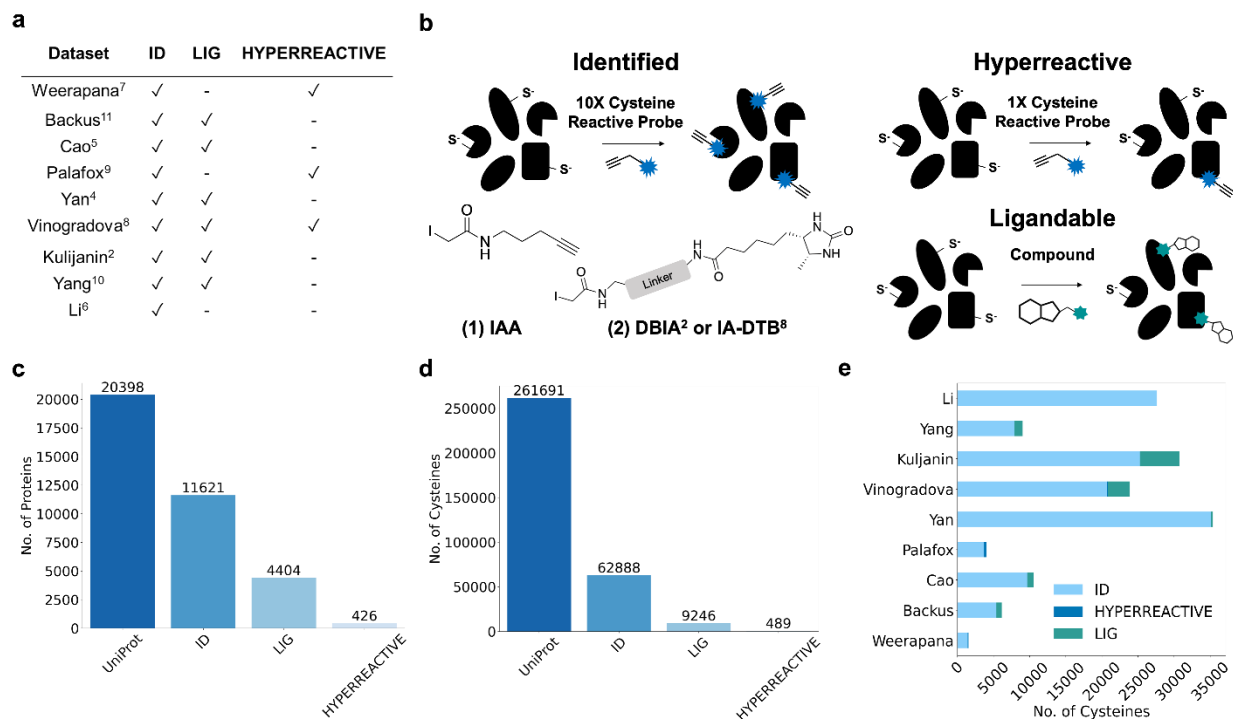
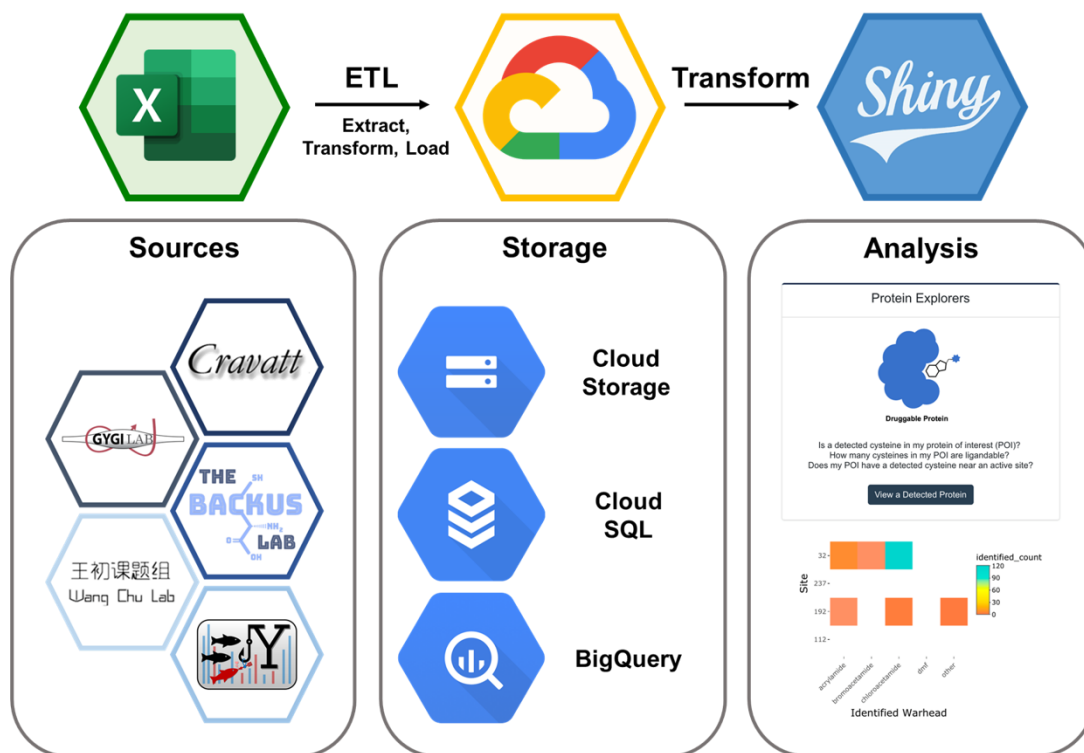


Figure 1. Dataset selection and curation for the creation of CysDB. (A) Table of all datasets used as input for CysDB, including which datasets were utilized in each chemoproteomic category (identified, hyperreactive and ligandable)^{2,4-11}. (B) General workflows for three categories of chemoproteomic methods included in CysDB that use iodoacetamide alkyne (**IAA**, **1**) or an iodoacetamide desthiobiotin reagent (**DBIA**² or **IA-DTB**⁸, **2**) to capture cysteines for: (i) high coverage identification of cysteine-containing peptides. (ii) quantitative profiling of intrinsic cysteine reactivity, and (iii) assaying cysteine ligandability using an electrophile of interest. (C-D) Quantification of the unique proteins (C) and cysteines (D) found in the Human UniProtKB/Swiss-Prot database, together with the identified, ligandable, and hyperreactive chemoproteomics subsets in CysDB. (E) Study-specific breakdown of total number of unique cysteines, including those that are identified as hyperreactive and ligandable. Data available in **Table S1**.

a



b

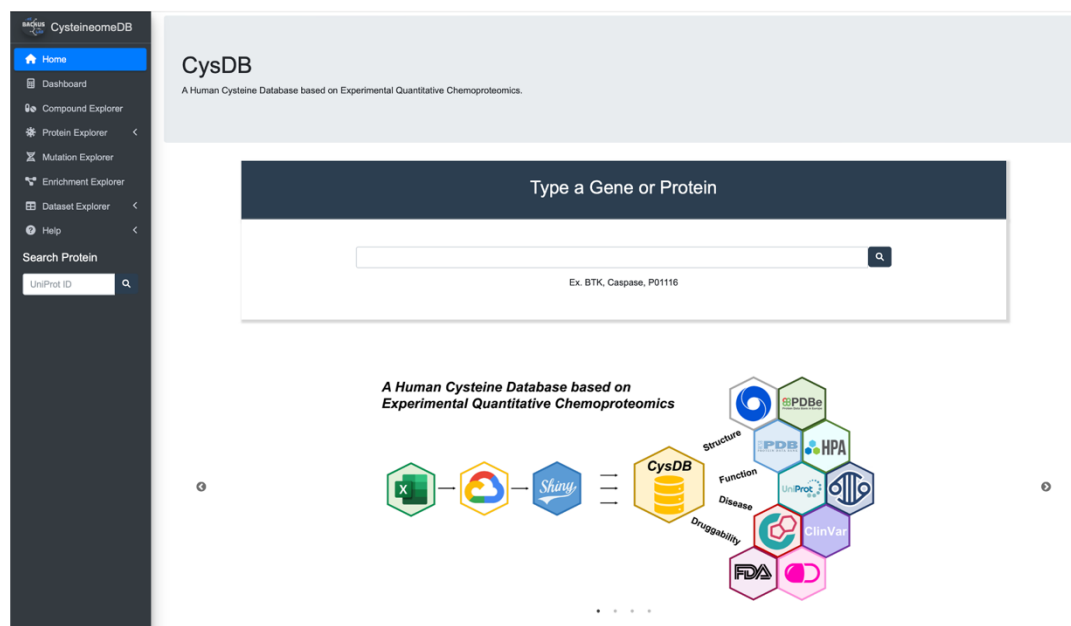


Figure 2. Workflow to generate CysDB SQL database. (A) Data extracted from nine datasets

(**Table S1**) was transformed and loaded into a MySQL relational database on the Google Cloud Platform. An accompanying front-end web interface was developed using RShiny to allow for remote-user querying of the SQL database. (B) Home page of the CysDB app publicly available at <https://backuslab.shinyapps.io/cysdb/>.

a

CysteineomeDB

Home

Dashboard

Compound Explorer

Protein Explorer

Activity

Structure

Function

Mutation Explorer

Enrichment Explorer

Dataset Explorer

Help

Search Protein

P78417

UniProtKB

AlphaFold

CysteineomeDB

Home

Dashboard

Compound Explorer

Protein Explorer

Activity

Structure

Function

Mutation Explorer

Enrichment Explorer

Dataset Explorer

Help

Search Protein

P78417

UniProtKB

AlphaFold

PDB

Ligandability Comparison Between Different Compounds

Select Datasets

Select Cell Lines

Select Range of Competition Ratios

backus_cravatt_ligandable, kunjanin_gygi_ligandable, vinogradov

MDA-MB-231, HCT116, Ramos, HEK293T, T Cell, PaTu-8988T, J...

0 2 4 6 8 10 12 14 16 18 20

Competition Ratio

resid

C192

C237

C32

C90

Individual Compounds Identifier

Compounds used in specific cell lines are used to create an individual 'Compound' identifier.

Ligandability Comparison of the Same Compound Used in Multiple Cell Lines

Compound Competition Ratios

resid

smiles

group_name

compound_name

competition_ratio

cell_line

datasetid

C112

C=CC(=O)N1CC2C(=O)C(C)C2C1

GROUP_ACRYL_107

ACRYL_515

1.00116

Ramos

yang_wang_ligandable

PDBs

Stage

Select

Label

Assoc. 3D Structures

Identifier

Id

method

chai

PDB

1EEM

X-ray

A=1

PDB

3LFL

X-ray

A/B

241

PDB

3VLN

X-ray

A=1

PDB

4ISO

X-ray

A=1

PDB

4YQM

X-ray

A/B

241

1-5 of 17 rows

Previous

1

2

3

4

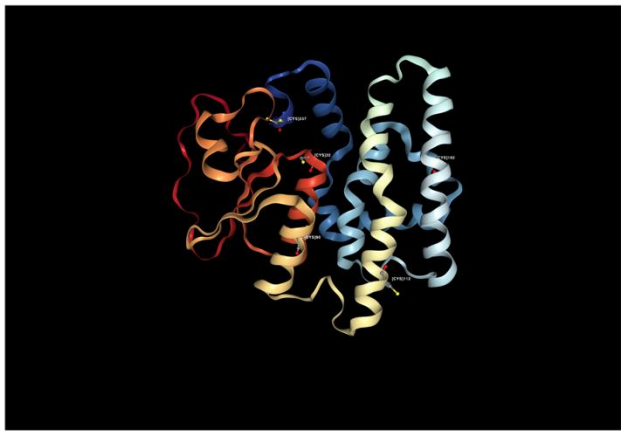
Next

Select PDB

SYVN

Selected 3D Structure

Snapshot



Protein Name

Glutathione S-transferase omega-1 (GSTO-1) (EC 2.5.1.18) (Glutathione S-transferase omega 1-1) (GSTO 1-1) (Glutathione-dependent dehydroascorbate reductase) (EC 1.8.5.1) (Monomethylarsonic acid reductase) (MMA(V) reductase) (EC 1.20.4.2) (S-(Phenacyl)glutathione reductase) (SPG-R)

Gene Name(s)

GSTO1;GSTTL28

Subcellular Location

Cytoplasm

Protein Function

Exhibits glutathione-dependent thiol transferase and dehydroascorbate reductase activities;Has S-(phenacyl)glutathione reductase activity;Has also glutathione S-transferase activity;Participates in the biotransformation of inorganic arsenic and reduces monomethylarsonic acid (MMA) and dimethylarsonic acid.

Reactome Pathways

Tissues

Diseases

GO

Keywords

Id

pathway name

R-HSA-156581

Methylation

R-HSA-156590

Glutathione conjugation

R-HSA-196836

Vitamin C (ascorbate) metabolism

R-HSA-8950505

Gene and protein expression by JAK-STAT signaling after Interleukin-12 stimulation

Protein Functional Terms Enrichment

Functional Enrichment Analysis

-log10(p-adj)

GO:0005575

GO:0005576

GO:0005577

GO:0005578

GO:0005579

GO:0005580

GO:0005581

GO:0005582

GO:0005583

GO:0005584

GO:0005585

GO:0005586

GO:0005587

GO:0005588

GO:0005589

GO:0005590

GO:0005591

GO:0005592

GO:0005593

GO:0005594

GO:0005595

GO:0005596

GO:0005597

GO:0005598

GO:0005599

GO:0005600

GO:0005601

GO:0005602

GO:0005603

GO:0005604

GO:0005605

GO:0005606

GO:0005607

GO:0005608

GO:0005609

GO:0005610

GO:0005611

GO:0005612

GO:0005613

GO:0005614

GO:0005615

GO:0005616

GO:0005617

GO:0005618

GO:0005619

GO:0005620

GO:0005621

GO:0005622

GO:0005623

GO:0005624

GO:0005625

GO:0005626

GO:0005627

GO:0005628

GO:0005629

GO:0005630

GO:0005631

GO:0005632

GO:0005633

GO:0005634

GO:0005635

GO:0005636

GO:0005637

GO:0005638

GO:0005639

GO:0005640

GO:0005641

GO:0005642

GO:0005643

GO:0005644

GO:0005645

GO:0005646

GO:0005647

GO:0005648

GO:0005649

GO:0005650

GO:0005651

GO:0005652

GO:0005653

GO:0005654

GO:0005655

GO:0005656

GO:0005657

GO:0005658

GO:0005659

GO:0005660

GO:0005661

GO:0005662

GO:0005663

GO:0005664

GO:0005665

GO:0005666

GO:0005667

GO:0005668

GO:0005669

GO:0005670

GO:0005671

GO:0005672

GO:0005673

GO:0005674

GO:0005675

GO:0005676

GO:0005677

GO:0005678

GO:0005679

GO:0005680

GO:0005681

GO:0005682

GO:0005683

GO:0005684

GO:0005685

GO:0005686

GO:0005687

GO:0005688

GO:0005689

GO:0005690

GO:0005691

GO:0005692

GO:0005693

GO:0005694

GO:0005695

GO:0005696

GO:0005697

GO:0005698

GO:0005699

GO:0005700

GO:0005701

GO:0005702

GO:0005703

GO:0005704

GO:0005705

GO:0005706

GO:0005707

GO:0005708

GO:0005709

GO:0005710

GO:0005711

GO:0005712

GO:0005713

GO:0005714

GO:0005715

GO:0005716

GO:0005717

GO:0005718

GO:0005719

GO:0005720

GO:0005721

GO:0005722

GO:0005723

GO:0005724

GO:0005725

GO:0005726

GO:0005727

GO:0005728

GO:0005729

GO:0005730

GO:0005731

GO:0005732

GO:0005733

GO:0005734

GO:0005735

GO:0005736

GO:0005737

GO:0005738

GO:0005739

GO:0005740

GO:0005741

GO:0005742

GO:0005743

GO:0005744

GO:0005745

GO:0005746

GO:0005747

GO:0005748

GO:0005749

GO:0005750

GO:0005751

GO:0005752

GO:0005753

GO:0005754

GO:0005755

GO:0005756

GO:0005757

GO:0005758

GO:0005759

GO:0005760

GO:0005761

GO:0005762

GO:0005763

GO:0005764

GO:0005765

GO:0005766

GO:0005767

GO:0005768

GO:0005769

GO:0005770

GO:0005771

GO:0005772

GO:0005773

GO:0005774

GO:0005775

GO:0005776

GO:0005777

GO:0005778

GO:0005779

GO:0005780

GO:0005781

GO:0005782

GO:0005783

GO:0005784

GO:0005785

GO:0005786

GO:0005787

GO:0005788

GO:0005789

GO:0005790

GO:0005791

GO:0005792

GO:0005793

GO:0005794

GO:0005795

GO:0005796

GO:0005797

GO:0005798

GO:0005799

GO:0005800

GO:0005801

GO:0005802

GO:0005803

GO:0005804

GO:0005805

GO:0005806

GO:0005807

GO:0005808

GO:0005809

GO:0005810

GO:0005811

GO:0005812

GO:0005813

GO:0005814

GO:0005815

GO:0005816

GO:0005817

GO:0005818

GO:0005819

GO:0005820

GO:0005821

GO:0005822

GO:0005823

GO:0005824

GO:0005825

GO:0005826

GO:0005827

GO:0005828

GO:0005829

GO:0005830

GO:0005831

GO:0005832

GO:0005833

GO:0005834

GO:0005835

GO:0005836

GO:0005837

GO:0005838

GO:0005839

GO:0005840

GO:0005841

GO:0005842

GO:0005843

GO:0005844

GO:0005845

GO:0005846

GO:0005847

GO:0005848

GO:0005849

GO:0005850

GO:0005851

GO:0005852

GO:0005853

GO:0005854

GO:0005855

GO:0005856

GO:0005857

GO:0005858

GO:0005859

GO:0005860

GO:0005861

GO:0005862

GO:0005863

GO:0005864

GO:0005865

GO:0005866

GO:0005867

GO:0005868

GO:0005869

GO:0005870

GO:0005871

GO:0005872

GO:0005873

GO:0005874

GO:0005875

GO:0005876

GO:0005877

GO:0005878

GO:0005879

GO:0005880

GO:0005881

GO:0005882

GO:0005883

GO:0005884

GO:0005885

GO:0005886

GO:0005887

GO:0005888

GO:0005889

GO:0005890

GO:0005891

GO:0005892

GO:0005893

GO:0005894

GO:0005895

GO:0005896

GO:0005897

GO:0005898

GO:0005899

GO:0005900

GO:0005901

GO:0005902

GO:0005903

GO:0005904

GO:0005905

GO:0005906

GO:0005907

GO:0005908

GO:0005909

GO:0005910

GO:0005911

GO:0005912

GO:0005913

GO:0005914

GO:0005915

GO:0005916

GO:0005917

GO:0005918

GO:0005919

GO:0005920

GO:0005921

GO:0005922

GO:0005923

GO:0005924

GO:0005925

GO:0005926

GO:0005927

GO:0005928

GO:0005929

GO:0005930

GO:0005931

GO:0005932

GO:0005933

GO:0005934

GO:0005935

GO:0005936

GO:0005937

GO:0005938

GO:0005939

GO:0005940

GO:0005941

GO:0005942

GO:0005943

GO:0005944

GO:0005945

GO:0005946

GO:0005947

GO:0005948

GO:0005949

GO:0005950

GO:0005951

GO:0005952

GO:0005953

GO:0005954

GO:0005955

GO:0005956

GO:0005957

GO:0005958

GO:0005959

GO:0005960

GO:0005961

GO:0005962

GO:0005963

GO:0005964

GO:0005965

GO:0005966

GO:0005967

GO:0005968

GO:0005969

GO:0005970

GO:0005971

GO:0005972

GO:0005973

GO:0005974

GO:0005975

GO:0005976

GO:0005977

GO:0005978

GO:0005979

GO:0005980

GO:0005981

GO:0005982

GO:0005983

GO:0005984

GO:0005985

GO:0005986

GO:0005987

GO:0005988

GO:0005989

GO:0005990

GO:0005991

GO:0005992

GO:0005993

GO:0005994

GO:0005995

GO:0005996

GO:0005997

GO:0005998

GO:0005999

GO:0006000

GO:0006001

GO:0006002

GO:0006003

GO:0006004

GO:0006005

GO:0006006

GO:0006007

GO:0006008

GO:0006009

GO:0006010

GO:0006011

GO:0006012

GO:0006013

GO:0006014

GO:0006015

GO:0006016

GO:0006017

GO:0006018

GO:0006019

GO:0006020

GO:0006021

GO:0006022

GO:0006023

GO:0006024

GO:0006025

GO:0006026

GO:0006027

GO:0006028

GO:0006029

GO:0006030

GO:0006031

GO:0006032

GO:0006033

GO:0006034

GO:0006035

GO:0006036

GO:0006037

GO:0006038

GO:0006039

GO:0006040

GO:0006041

GO:0006042

GO:0006043

GO:0006044

GO:0006045

GO:0006046

GO:0006047

GO:0006048

GO:0006049

GO:0006050

GO:0006051

GO:0006052

GO:0006053

GO:0006054

GO:0006055

GO:0006056

GO:0006057

GO:0006058

GO:0006059

GO:0006060

GO:0006061

GO:0006062

GO:0006063

GO:0006064

GO:0006065

GO:0006066

GO:0006067

GO:0006068

GO:0006069

GO:0006070

GO:0006071

GO:0006072

GO:0006073

GO:0006074

GO:0006075

GO:0006076

GO:0006077

GO:0006078

GO:0006079

GO:0006080

GO:0006081

GO:0006082

GO:0006083

GO:0006084

GO:0006085

GO:0006086

GO:0006087

GO:0006088

GO:0006089

GO:0006090

GO:0006091

GO:0006092

GO:0006093

GO:0006094

GO:0006095

GO:0006096

GO:0006097

GO:0006098

GO:0006099

GO:0006100

GO:0006101

GO:0006102

GO:0006103

GO:0006104

GO:0006105

GO:0006106

GO:0006107

GO:0006108

GO:0006109

GO:0006110

GO:0006111

GO:0006112

GO:0006113

GO:0006114

GO:0006115

GO:0006116

GO:0006117

GO:0006118

GO:0006119

GO:0006120

GO:0006121

GO:0006122

GO:0006123

GO:0006124

GO:0006125

GO:0006126

GO:0006127

GO:0006128

GO:0006129

GO:0006130

GO:0006131

GO:0006132

GO:0006133

GO:0006134

GO:0006135

GO:0006136

GO:0006137

GO:0006138

GO:0006139

GO:0006140

GO:0006141

GO:0006142

GO:0006143

GO:0006144

GO:0006145

GO:0006146

GO:0006147

GO:0006148

GO:0006149

GO:0006150

GO:0006151

GO:0006152

GO:0006153

GO:0006154

GO:0006155

GO:0006156

GO:0006157

GO:0006158

GO:0006159

GO:0006160

GO:0006161

GO:0006162

GO:0006163

GO:0006164

GO:0006165

GO:0006166

GO:0006167

GO:0006168

GO:0006169

GO:0006170

GO:0006171

GO:0006172

GO:0006173

GO:0006174

GO:0006175

GO:0006176

GO:0006177

GO:0006178

GO:0006179

GO:0006180

GO:0006181

GO:0006182

GO:0006183

GO:0006184

GO:0006185

GO:0006186

GO:0006187

GO:0006188

GO:0006189

GO:0006190

GO:0006191

GO:0006192

GO:0006193

GO:0006194

GO:0006195

GO:0006196

GO:0006197

GO:0006198

GO:0006199

GO:0006200

GO:0006201

GO:0006202

GO:0006203

GO:0006204

GO:0006205

GO:0006206

GO:0006207

GO:0006208

GO:0006209

GO:0006210

GO:0006211

GO:0006212

GO:0006213

GO:0006214

GO:0006215

GO:0006216

GO:0006217

GO:0006218

GO:0006219

GO:0006220

GO:0006221

GO:0006222

GO:0006223

GO:0006224

GO:0006225

GO:0006226

GO:0006227

GO:0006228

GO:0006229

GO:0006230

GO:0006231

GO:0006232

GO:0006233

GO:0006234

GO:0006235

GO:0006236

GO:0006237

GO:0006238

GO:0006239

GO:0006240

GO:0006241

GO:0006242

GO:0006243

GO:0006244

GO:0006245

GO:0006246

GO:0006247

GO:0006248

GO:0006249

GO:0006250

GO:0006251

GO:0006252

GO:0006253

GO:0006254

GO:0006255

GO:0006256

GO:0006257

GO:0006258

GO:0006259

GO:0006260

GO:0006261

GO:0006262

GO:0006263

GO:0006264

GO:0006265

GO:0006266

GO:0006267

GO:0006268

GO:0006269

GO:0006270

GO:0006271

GO:0006272

GO:0006273

GO:0006274

GO:0006275

GO:0006276

GO:0006277

GO:0006278

GO:0006279

GO:0006280

GO:0006281

GO:0006282

GO:0006283

GO:0006284

GO:0006285

GO:0006286

GO:0006287

GO:0006288

GO:0006289

GO:0006290

GO:0006291

GO:0006292

GO:0006293

GO:0006294

GO:0006295

GO:0006296

GO:0006297

GO:0006298

GO:0006299

GO:0006300

GO:0006301

GO:0006302

GO:0006303

GO:0006304

GO:0006305

GO:0006306

GO:0006307

GO:0006308

GO:0006309

GO:0006310

GO:0006311

GO:0006312

GO:0006313

GO:0006314

GO:0006315

GO:0006316

GO:0006317

GO:0006318

GO:0006319

GO:0006320

GO:0006321

GO:0006322

GO:0006323

GO:0006324

GO:0006325

GO:0006326

GO:0006327

GO:0006328

GO:0006329

GO:0006330

GO:0006331

GO:0006332

GO:0006333

GO:0006334

GO:0006335

GO:0006336

GO:0006337

GO:0006338

GO:0006339

GO:0006340

GO:0006341

GO:0006342

GO:0006343

GO:0

CysteineomeDB

Home

Dashboard

Compound Explorer

Protein Explorer

Activity

Structure

Function

Mutation Explorer

Enrichment Explorer

Dataset Explorer

Help

Search Protein

P78417

GSTO1_HUMAN

UniProtKB

AlphaFold

PDB

5 CysDB Cysteines

0 Pathogenic Missense Variants

0 Cancer Census Genes

Protein Schematic with Labeled Variants

GSTO1_HUMAN

Amino acid number

Scheme Options

Select Features to be Displayed

☒ CHAIN
 ☐ DOMAIN
 ☐ ACT_SITE
 ☐ BINDING
 ☐ MOD_RES
 ☐ MUTAGEN
 ☐ HELIX

Protein Features

description	begin	end	alternativeSequence
in dbSNP:rs45529437	32	32	Y
in dbSNP:rs11509436	86	86	C
in allele GSTO1*1C; no effect on protein stability: dbSNP:rs4925	140	140	D

Enrichment Options

Select Range of Reactivity Ratios

0 2 4 6 8 10 12 14 16 18 20

Select Enrichment Library

GO_Biological_Process_2021

Submit

Enrichment of Reactive Proteins Results

Show 2 entries

Search:

Term	Overlap	P-value	Adjusted P-value	Odds Ratio	Combined Score
gene expression (GO:0010467)	205/356	0	0	5.611	703.765
rRNA processing (GO:0006364)	129/173	0	0	11.962	1483.708

Showing 1 to 2 of 5,232 entries

Previous 1 2 3 4 5 ... 2,616 Next

Download

Enrichment Analysis

GO_Biological_Process_2021

P-value

5.575346e-38

1.932760e-38

1.285173e-38

6.415966e-39

3.382599e-45

Compound Physicochemical Properties

ACRYL_1

Q

Compound Structure

Passes Lipinski Filter

yes

Passes Ghose Filter

Passes Rule of Three Filter

Molecular Weight

361.01

Heavy Atoms

18

Hbond Acceptors

1

Hbond Donors

0

Rotable Bonds

3

logP

4.30

Polar Surface Area

20.31

Molar Refractivity

77.93

Grouped Compound Protein Targets

Select Group Compound:

GROUP_ACRYL_1

0 Group Minimum Ratio

20 Group Maximum Ratio

6014 Labeled Proteins

5 Proteins with (R>= 10)

Figure 3. CysDB outputs based on protein (A), disease (B), dataset (C) and cysteine-reactive compound wise queries (D). (A) Users can search for a protein of interest (POI) in the search bar on the protein page. Centered on the activity tab is a ‘site map,’ indicating which cysteines have been identified, liganded or hyperreactive by chemoproteomics. In addition, the activity tab allows users to assess the potential druggability of their POI through small-molecule binding annotations and heatmaps for quantitative chemoproteomic measures of hyperreactivity and ligandability. For a comprehensive view of the structural environment surrounding the chemoproteomic detected cysteines, publicly available 3D crystal structures are displayed in the structure tab. Users can choose which structure is shown and add customized labels. By clicking the function tab, one can view general information on the POI, including subcellular locations, functional pathways and GO/KEGG terms. (B) The disease-relevance of a POI can be explored through the mutation page. Proximity of chemoproteomic detected cysteines, annotated small-molecule binders and variants of ranging clinical significance are visualized on a one-dimensional schematic of a protein sequence. Chemoproteomic cysteines are colored in gold for identified, pink for ligandable and orange for hyperreactive, while the remaining points are variant positions. (C) Users can specify subsets of data available in CysDB, such as by compound chemotype or ranges of reactivity ratio, for pathway, ontology and disease enrichment analyses. From these dataset wise queries on the enrichment page, a user can then download their results as a CSV formatted table or a bar graph as an image. (D) Chemical structures and calculated ‘drug-likeness’ properties of compounds used to ligand cysteines in CysDB can be accessed from the dropdown menu in the compound page.

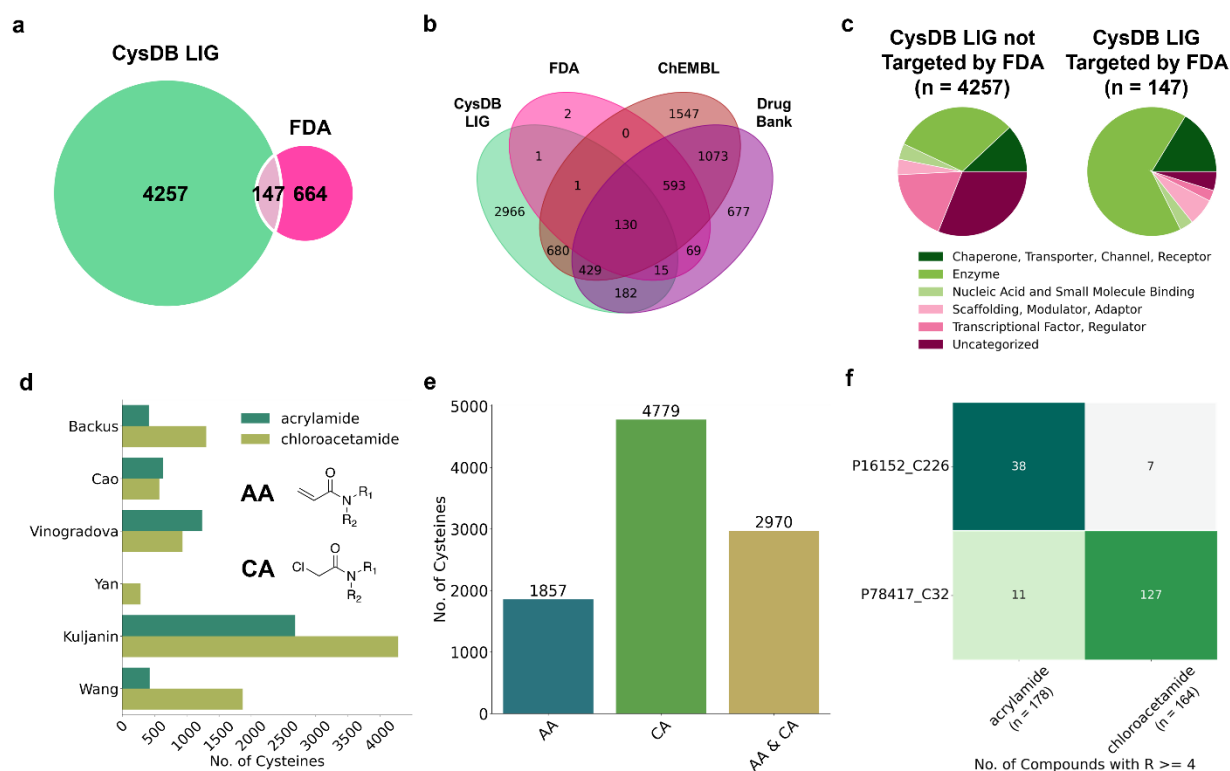


Figure 4. Cysteines with available ligandability data. (A) Overlap between CysDB ligandable (LIG) proteins and proteins targeted by FDA approved drugs. (B) Overlap between CysDB LIG proteins, proteins targeted by FDA approved drugs, small molecules in DrugBank and ChEMBL. (C) Distributions of protein functions for CysDB LIG proteins not targeted by FDA and CysDB LIG proteins targeted by FDA. (D) Grouped bar graph showing the number of unique ligandable cysteines targeted by acrylamides or chloroacetamide for each dataset ($R \geq 4$ for at least one compound). (E) Bar graph of the overall number of unique cysteines targeted by acrylamides or chloroacetamide. (F) Number of unique SMILES strings with an acrylamide and chloroacetamide moiety (based on the 'Group Compound Identifier'), compounds with a ratio ≥ 4 for protein carbonyl reductase (CBR1, UniProtKB: P16512) and protein glutathione s-transferase omega-1 (GSTO1, UniProtKB: P78417). Data available in **Table S2**.

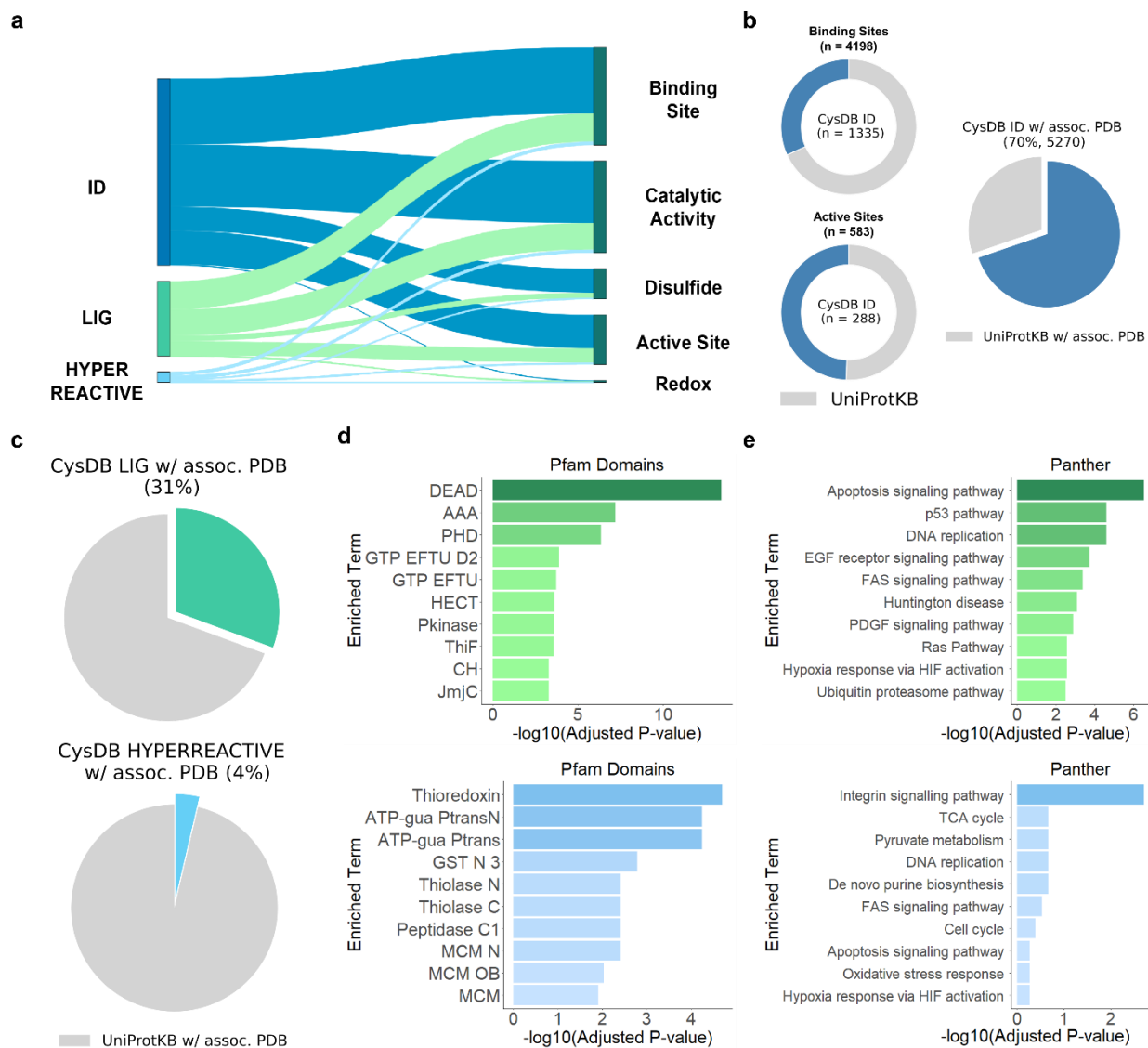


Figure 5. Cysteines with available functional and structural annotations. (A) CysDB identified, ligandable and hyperreactive proteins with annotated active sites, binding sites, catalytic activity, disulfide bonds and redox potentials. (B) Distribution of identified cysteines in CysDB ID annotated as cysteine-specific binding sites or active sites (left). The total number of cysteines in UniProtKB annotated as binding or active sites are shown in gray. Percentage of proteins associated with a PDB structure and contain an identified cysteine. (C) Percentage of proteins associated with a PDB structure and contain a ligandable (CysDB LIG) or

hyperreactive (CysDB HYPERREACTIVE) cysteine. (D) Top-10 enriched protein domains from Pfam-term enrichment analysis of liganded (green) and hyperreactive (light blue) proteins. (E) Top-10 enriched pathways from Panther-term enrichment analysis of liganded (green) and hyperreactive (light blue) proteins. Data available in **Table S3**.

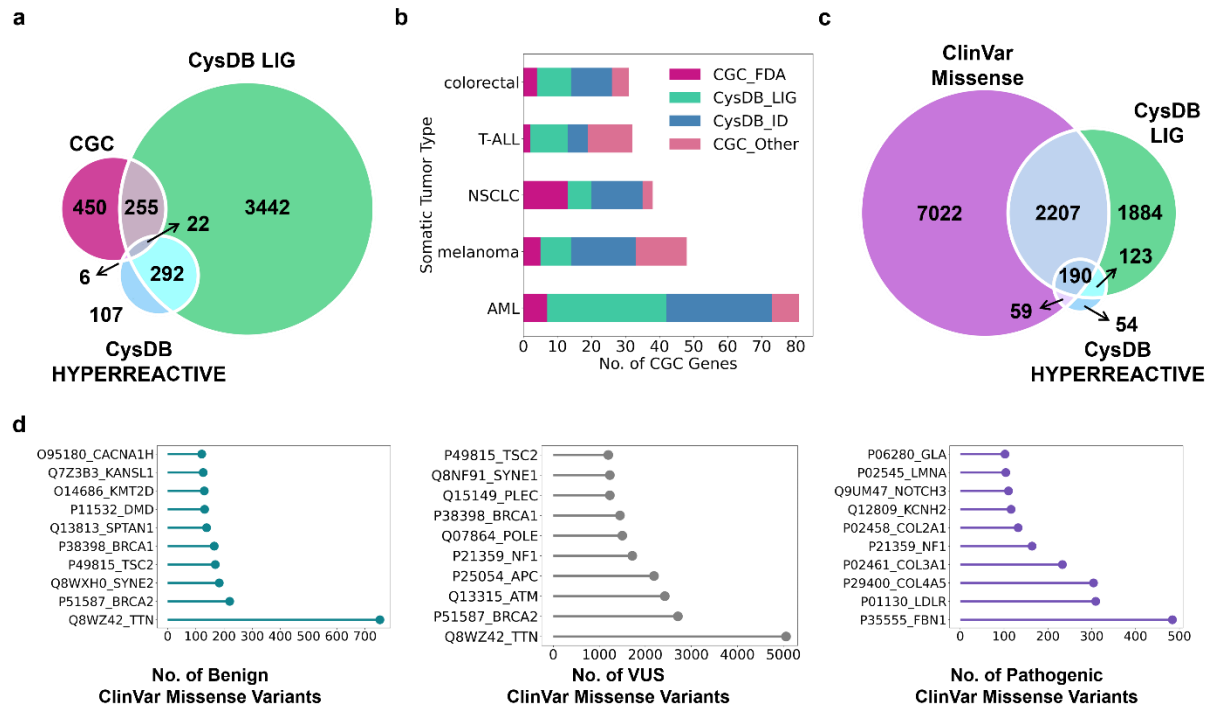


Figure 6. Assessment of the scope of disease-relevant proteins contained in CysDB of biologically relevant proteins using cysteine chemoproteomics. (A) Overlap between genes associated with cancer by the Cancer Gene Census (CGC), genes associated with CysDB ligandable proteins and genes associated with CysDB hyperreactive proteins. (B) For the five most abundant tumor types in CGC, the number of CGC genes targeted by FDA approved drugs (CGC_FDA), non-FDA targeted CGC genes identified in CysDB (CysDB_ID), non-FDA targeted CGC genes liganded in CysDB (CysDB_LIG) and non-FDA targeted CGC genes not identified in CysDB (CGC_Other). (C) Overlap between unique proteins associated with ClinVar genes

containing missense variants (9,951 genes mapped to 9,478 proteins), CysDB ligandable proteins and CysDB hyperreactive proteins. (D) Top ten CysDB identified proteins with the highest number of benign missense variants (teal), missense variants of unknown significance (VUS) (gray) and pathogenic missense variants (purple). Data available in **Table S4**.