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*}

- 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.
- 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 ~99% 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 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 cysteineome.

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, <1% 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 peptides using label free quantification⁴⁻⁶. (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 this year alone^{6,10,14}.

The availability of easily searchable cysteine databases—including Oximouse¹, the Ligandable Cysteine Database, and previously reported Cysteinome¹⁵—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)²⁵⁻³¹, 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 facilitate 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.

To enable the creation of the CysDB, our first step was to curate a set of publicly available datasets. With the overarching goal of establishing a high coverage and highly curated database to facilitate 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) and (3) assaying cysteine ligandability (**Figure 1A**). We collected a total of nine datasets that fulfilled our criteria (**Figure 1B** for all datasets used)^{2,4-11}.

Notably, all of 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., **DBIA² or IA-DTB⁸**) 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 stratify 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 10 μ M and 100 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 facilitate the production of 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 publication's 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 UniProt 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 UniProt releases used in dataset search, as observed in our prior study⁹. For multi-cysteine-containing peptides with probe labeling detected at more than one cysteine, we generated separate identifiers for each cysteine. 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. Thereby, limiting the potential impact of cell-state dependent differences of cysteine reactivity as a potential confounder to our downstream analyses. 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 & 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 (99.3%) 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 cyanoacrylamides, fumarates, and activated esters—while activated esters are primarily lysine reactive our prior data indicates that they do also exhibit cysteine-reactivity^{32,33}.

All datasets included in our database relied on competition ratio cutoffs for what defines a cysteine as 'ligandable.' Peptides included in the aggregate dataset (those used for further bioinformatics and statistical analyses) were required to have been quantified in 3 experiments. Cysteines were categorized as liganded if they had at least two ratios $R \ge 4$ (hit fragments) and one ratio between 0.5 and 2 (control fragments). When processing the ligandability data for each dataset, we observed manuscript-specific differences in the requirements for designating a cysteine as ligandable. Exemplifying these differences, 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. Additionally, we observed non-universal data filtration strategies, including, for example, a requirement for an elevated ligandability ratio for at least two unique compounds. In contrast, other studies were more permissive and included cysteines with only single compound ligandability^{2,6,8}. 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. Otherwise, despite the differences in defining ligandability, we opted to retain all remaining liganded cysteines to accurately represent each study's reported findings. 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 processed the raw 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 (10 μ M and 100 μ M), where a quantitative isoTOP-ABPP ratio (R_[high]:[low]) reflects the differences in signal intensities between the 100 µM and 10 µM treated proteomes. Highly reactive cysteines, termed 'hyperreactive' residues, were 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 $R_{100:10} >= 2$ 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 S5**). A key outcome of the dataset aggregation required to build CysDB is an effective doubling of the size of the identified cysteineome. Collectively across all studies analyzed in CysDB, ~25% of all cysteines found on 57% of proteins in UniProt 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, our next step was to construct the CysDB database and web user interface outlined in **Figure 2A**. Raw data from prioritized studies^{2,4-11} were pre-processed into a standardized input format for SQL integration (See **Table S1** for example data format and required information for future data integration to CysDB). Processed data from these selected sources was ingested and transformed 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 UniProt, COSMIC, ClinVar and PDB^{25-28,31}.

One challenge we faced during our processing of the raw data 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)³⁴ 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 UniProt accession⁹. In CysDB, we found that 16 UniProt entries were associated with multiple gene names. To address this limitation, we opted to construct CysDB using UniProKB accession numbers.

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 UniProt ID returns the 'Protein Section,' which is further broken up into three separate tabs detailing function, activity, and structure. The function tab reports functional annotations for the POI generated from UniProt, Gene Ontology (GO), Reactome, and STRING^{25,35-37}, as well as a 'site map' indicating whether any cysteines in the POI that are hyperreactive or ligandable. The activity tab provides further stratification of cysteine hyperreactivity and ligandability, including the measured reactivity and competition ratios and the structures of all compounds that ligand the POI. Lastly, the structure tab provides the user with an easily accessible mechanism to visualize the three-dimensional protein microenvironment of chemoproteomic detected cysteines, including for structures reported in the PDB.

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 variant 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 facilitate identification of clinically relevant protein regions containing ligandable and hyperreactive cysteines, the Disease Section of CysDB also provides the clinical significance for

variants as reported by Clinvar²⁷, 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 package³⁸⁻³⁹ (**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. Included in this section are details of each compound used in the ligandability experiments, the CysDB compound abbreviation, corresponding publication abbreviation, and dataset. To facilitate future studies, this information is also provided as an easily downloadable table. Selection of individual compounds using the provided drop down menu 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**)⁴⁰⁻⁴⁵.

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. To further facilitate future studies that harness the CysDB datasets, the Statistics Section also provides downloadable versions of the aggregated and individual datasets

used to build CysDB—these datasets are also provided as supporting tables alongside this manuscript (**Table S1**).

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

With the CysDB database established, our next step was to further stratify and parse the data available in CysDB with the overarching goals of showcasing the enabling features built into CysDB and facilitating 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. We find that less than 1% of all human proteins in UniProt 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 revealed 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 UniProt 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, revealed a marked 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 notable enrichment for 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 stratified 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 labeling⁴⁶⁻⁴⁹. 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., alpha-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 structureactivity relationships (SAR) for different cysteine-reactive electrophiles, we next quantified the number of cysteines detected as labeled by individual electrophile chemotypes (Figure 4D, Figure S8 and Figure S9)^{2,50-55}. 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 (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 a marked 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's the highly ligandable cysteine (Cys 32) exhibits strong preference for reacting with chloroacetamide (CA)-substituted compounds (1 to 29.5 in

favor of CA electrophiles). In contrast, cysteine 226 of CBR1 shows marked acrylamide (AA) bias (15.5 to 1 in preference of AA warheads).

(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 facilitate discovery of likely functional and disease-relevant cysteines, CysDB includes metrics of functionality from UniProt, 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 UniProt annotations to determine which CysDB proteins had functional annotations of the following active sites, binding sites, catalytic activity, disulfide bonds and redox potentials. Analysis revealed 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 6 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 UniProt site annotations for residue positions. This analysis revealed 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 cysteines found in binding sites and 288 (49%) of all known cysteine active

sites (**Figure S13**). Out of the 4,198 cysteine specific binding sites, 178 of them have been liganded by a compound in CysDB. In addition, 90 out of the 583 cysteine active sites have been liganded by a compound in CysDB.

Next, we extended this analysis to look for cysteines '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, 574 ligandable and 46 hyperreactive CysDB cysteines are near binding sites (**Figure S14**) and 154 ligandable and 53 hyperreactive CysDB cysteines are near active sites (**Figure S15**). Consistent with measures of cysteine hyper-reactivity as providing a useful surrogate for cysteine functionality, we observed a marked increase for active sites in the hyperreactive subset compared with the ligandable subset (**Figure 5B**).

As the Uniprot 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 (70%) of CysDB ID proteins are associated with an available PDB. Of these, 2,314 (31%) contain one or more ligandable cysteines and 279 feature at least one hyperreactive cysteine (**Figure 5C**). To establish 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 UniProt and PDB residue numbers, residue to protein sequence numbering was mapped using SIFT annotations⁵⁶ (**Figure S16**). This systematic analysis revealed that 4,733 (14%) of CysDB identified cysteines are resolved in a corresponding crystal structure. Further stratification of this dataset revealed that 479 CysDB cysteines are proximal (within 10 Angstroms) to active site residues in 3D space (**Figure S19** and **Figure S20**). To facilitate 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)⁶⁰ domains elucidated a 30-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)⁵⁸. 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 medulloblastoma⁵⁹⁻⁶². 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 Panther⁶³ 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 marked enrichment indicates untapped opportunities for the development of probes targeting cell death⁶⁴⁻⁶⁵. The hyperreactive cysteine-containing protein by contrast was markedly 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 formation⁶⁶⁻⁶⁷.

(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 next assessed the human disease relevance of the CysDB proteins. Restricting our analysis to the ligandable and hyperreactive subsets, we next assessed which phenotypes were associated with CysDB proteins. Using disease annotations from the Online Mendelian Inheritance in Man (OMIM)⁶⁸ knowledge base, ligandable cysteine-containing proteins showed a ~1.5 fold-enrichment for terms related to a broad range of cancers, including colorectal, breast and leukemia. The hyperreactive cysteinecontaining protein subset was enriched for terms associated with immune-relevant diseases, specifically those affecting the lymphatic system (Figure S25). We next assessed how many CysDB proteins are annotated as cancer driving genes, as assessed by the Cancer Gene Census (CGC)²⁶. 77% of the proteins associated with CGC genes have been identified by CysDB (584/756) (Figure S28). 38% of CGC proteins are annotated as ligandable in CysDB, indicating untapped opportunities for the development of tailored therapies targeting driver mutations (Figure 6A). These results compare favorably to the 11% of proteins associated with cancer driving genes that have been targeted by FDA approved small molecules (Figure S29 and Table **S2**). We observed a marked 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 ClinVar²⁷ 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). 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 mutations²⁶. 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 S30). The proteins with the highest number of pathogenic variants are Fibrillin-1 (FBN1, UniProt: P35555) and Low-density lipoprotein receptor (LDLR, UniProt: 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 domains⁷⁰. Additionally, LDLR contains cysteinerich 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 disease⁷⁰⁻⁷². 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 Sotorasib73-76.

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 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 simplest 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 facilitate 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 log2(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 group⁷⁷.

Showcasing the utility of our efforts to generate CysDB, we find that by combining datasets generated across multiple cell lines and using different labeling reagents significantly increased our coverage of the cysteinome. We observed marked differences in cysteines identified in proteomes derived from different cell lines (Figure S5). 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. However, not only did cell line selection impact our number of identified cysteines, but also the hyperreactivity and ligandability of individual cysteines. In its current iteration, CysDB provides a low-throughput mechanism to assess reproducible ligandability of cysteines across studies. However, the absence of shared compounds used across multiple studies has limited reproducibility analysis at the level of specific compounds. The marked bias towards chemoproteomic analysis of chloroacetamide and acrylamides points to largely untapped opportunities in expanding the scope of the ligandable cysteinome through assaying additional classes of electrophiles. 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 facilitate efforts in assessing the reproducibility of ligandable ratios across studies.

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 marked 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 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 UniProt 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 sequences²⁵. Our use of UniProt accession numbers for query limits the potential for mismapping or multimapping during CysDB database search. Given the utility of a gene-centric search, we plan to incorporate such identifiers in future iterations of CysDB focused on facilitating future proteogenomic analysis.

An ongoing goal of CysDB is to facilitate expanding the scope of the ligandable and potentially druggable cysteineome, 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 future iterations of CysDB, alongside generalized strategies to merge the diverse data formats generated by these studies. Looking to the future, 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 Pride⁷⁸.

METHODS

Datasets used

Data sources	Release Date	URL		
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/c ensus		
ClinVar	2209-release	https://cancer.sanger.ac.uk/c ensus		
Human Protein Atlas (HPA)	Version 21.1	https://www.proteinatlas.org		
Enrichr	Accessed Sept. 2022	https://maayanlab.cloud/Enric hr/		
Enrichr Panther	2016	http://www.pantherdb.org/pat hway/		
Enrichr Pfam Domains	2019	https://pfam.xfam.org/		
Enrichr OMIM Disease		https://www.omim.org/downlo ads		

Data Collection and Processing

Chemoproteomics data was collected from publicly accessible supplementary tables of previous literature^{2,4-11}. Columns were parsed for UniProt protein identifiers and locations of the corresponding modified cysteine amino acid numbers to create a new identifier for CysDB: UniProtKBID_CYS#. For peptides modified at multiple cysteines, new identifiers were made for each cysteine position. 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**, **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_#. Publication names for each compound and CysDB names are provided in **Tables S2**.

Peptide Mapping to UniProt Identifiers

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 UniProt human fasta reference file, as this release is the only version saved in the UniProt archive for future mapping. Cysteines from unmatched peptides were removed prior to subsequent analyses.

Gene to Protein Identifier Mapping

Cancer Gene Census (CGC) website reports were downloaded Sept. 2022 and mapped to CysDB data using UniProt 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 UniProt codes (2209-release).

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 'druglikeness' 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. Finally, the warhead table holds chemotype classifications for each compound. The five chemotype classifications were as follows: acrylamide, bromoacetamide, chloroacetamide, dimethyl fumarate and other.

Web Server

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 (<u>https://biit.cs.ut.ee/gprofiler/gost</u>). All plots are generated with the ggplot2 and plotly (<u>https://plotly.com/r/</u>) R libraries.

Dataset Addition to CysDB Guidelines

Email submission materials to <u>cysteineomedb@gmail.com</u> 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 UniProt 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 >= 4 for at least two compounds). Table templates and additional information for submission requests can be found in **Table S1**.

Data and Code Availability

The data set and source code are available at https://github.com/Imboat/cysdb

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

- Xiao, H., Jedrychowski, M. P., Schweppe, D. K., Huttlin, E. L., Yu, Q., Heppner, D. E., ...
 & Chouchani, E. T. (2020). A quantitative tissue-specific landscape of protein redox regulation during aging. *Cell*, *180*(5), 968-983.
- Kuljanin, M., Mitchell, D. C., Schweppe, D. K., Gikandi, A. S., Nusinow, D. P., Bulloch, N. J., Vinogradova, E. V., Wilson, D. L., Kool, E. T., Mancias, J. D., Cravatt, B. F., & Gygi, S. P. (2021). Reimagining high-throughput profiling of reactive cysteines for cell-based screening of large electrophile libraries. *Nature Biotechnology*, *39*(5), 630–641.
- Müller, S., Ackloo, S., Al Chawaf, A., Al-Lazikani, B., Antolin, A., Baell, J. B., ... & Arrowsmith, C. H. (2022). Target 2035–update on the quest for a probe for every protein. *RSC Medicinal Chemistry*, *13*(1), 13-21.
- Yan, T., Desai, H. S., Boatner, L. M., Yen, S. L., Cao, J., Palafox, M. F., Jami-Alahmadi, Y., & Backus, K. (2021). SP3-FAIMS chemoproteomics for high coverage profiling of the human cysteinome. *ChemBioChem*.
- Cao, J., Boatner, L. M., Desai, H. S., Burton, N. R., Armenta, E., Chan, N. J., Castellón, J. O., & Backus, K. M. (2021). Multiplexed CuAAC Suzuki-Miyaura Labeling for Tandem Activity-Based Chemoproteomic Profiling. *Analytical Chemistry*, *93*(4), 2610–2618. https://doi.org/10.1021/acs.analchem.0c04726
- Li, Z., Liu, K., Xu, P., & Yang, J. (2022). Benchmarking Cleavable Biotin Tags for Peptide-Centric Chemoproteomics. *Journal of Proteome Research*, *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. (2010). Quantitative reactivity profiling predicts functional cysteines in proteomes. *Nature*, *468*(7325), 790–797. https://doi.org/10.1038/nature09472

- 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. (2020). An Activity-Guided Map of Electrophile-Cysteine Interactions in Primary Human T Cells. *Cell*, *182*(4), 1009-1026.e29.
- Palafox, M. F., Desai, H. S., Arboleda, V. A., & Backus, K. M. (2021). From chemoproteomic-detected amino acids to genomic coordinates: insights into precise multiomic data integration. *Molecular Systems Biology*, *17*(2). https://doi.org/10.15252/msb.20209840
- Yang, F., Jia, G., Guo, J., Liu, Y., & Wang, C. (2022). Quantitative Chemoproteomic Profiling with Data-Independent Acquisition-Based Mass Spectrometry. *Journal of the American Chemical Society*, 144(2), 901–911. https://doi.org/10.1021/jacs.1c11053
- Backus, K. M., Correia, B. E., Lum, K. M., Forli, S., Horning, B. D., González-Páez, G. E., Chatterjee, S., Lanning, B. R., Teijaro, J. R., Olson, A. J., Wolan, D. W., & Cravatt, B. F. (2016). Proteome-wide covalent ligand discovery in native biological systems. *Nature*, 534(7608), 570–574.
- Bar-Peled, L., Kemper, E. K., Suciu, R. M., Vinogradova, E. V., Backus, K. M., Horning,
 B. D., ... & Cravatt, B. F. (2017). Chemical proteomics identifies druggable vulnerabilities in a genetically defined cancer. *Cell*, *171*(3), 696-709.
- Backus, K. M. (2018). Applications of reactive cysteine profiling. *Activity-Based Protein Profiling*, 375-417.
- Abegg, D., Frei, R., Cerato, L., Prasad Hari, D., Wang, C., Waser, J., & Adibekian, A. (2015). Proteome-wide profiling of targets of cysteine reactive small molecules by using

ethynyl benziodoxolone reagents. Angewandte Chemie, 127(37), 11002-11007.

- Kulkarni, R. A., Bak, D. W., Wei, D., Bergholtz, S. E., Briney, C. A., Shrimp, J. H., ... & Meier, J. L. (2019). A chemoproteomic portrait of the oncometabolite fumarate. *Nature chemical biology*, *15*(4), 391-400.
- Grossman, E. A., Ward, C. C., Spradlin, J. N., Bateman, L. A., Huffman, T. R., Miyamoto,
 D. K., ... & Nomura, D. K. (2017). Covalent ligand discovery against druggable hotspots targeted by anti-cancer natural products. *Cell chemical biology*, *24*(11), 1368-1376.
- Tian, C., Sun, R., Liu, K., Fu, L., Liu, X., Zhou, W., ... & Yang, J. (2017). Multiplexed thiol reactivity profiling for target discovery of electrophilic natural products. *Cell Chemical Biology*, *24*(11), 1416-1427.
- Wang, C., Weerapana, E., Blewett, M. M., & Cravatt, B. F. (2014). A chemoproteomic platform to quantitatively map targets of lipid-derived electrophiles. *Nature methods*, *11*(1), 79-85.
- Abegg, D., Tomanik, M., Qiu, N., Pechalrieu, D., Shuster, A., Commare, B., ... & Adibekian,
 A. (2021). Chemoproteomic Profiling by Cysteine Fluoroalkylation Reveals Myrocin G as an Inhibitor of the Nonhomologous End Joining DNA Repair Pathway. *Journal of the American Chemical Society*, 143(48), 20332-20342.
- Fu, L., Li, Z., Liu, K., Tian, C., He, J., He, J., He, F., Xu, P., & Yang, J. (2020). A quantitative thiol reactivity profiling platform to analyze redox and electrophile reactive cysteine proteomes. *Nature Protocols*, *15*(9), 2891–2919. https://doi.org/10.1038/s41596-020-0352-2
- Desai, H. S., Yan, T., Yu, F., Sun, A. W., Villanueva, M., Nesvizhskii, A. I., & Backus, K. M. (2022). SP3-Enabled Rapid and High Coverage Chemoproteomic Identification of Cell-State–Dependent Redox-Sensitive Cysteines. *Molecular and Cellular Proteomics*, *21*(4), 100218.
- 22. Shi, Y., Fu, L., Yang, J., & Carroll, K. S. (2021). Wittig reagents for chemoselective sulfenic

acid ligation enables global site stoichiometry analysis and redox-controlled mitochondrial targeting. *Nature Chemistry*, *13*(11), 1140-1150.

- Mnatsakanyan, R., Markoutsa, S., Walbrunn, K., Roos, A., Verhelst, S. H., & Zahedi, R. P. (2019). Proteome-wide detection of S-nitrosylation targets and motifs using bioorthogonal cleavable-linker-based enrichment and switch technique. *Nature communications*, *10*(1), 1-12.
- 24. Wu, S., Luo (Howard), H., Wang, H., Zhao, W., Hu, Q., & Yang, Y. (2016). Cysteinome:
 The first comprehensive database for proteins with targetable cysteine and their covalent inhibitors. *Biochemical and Biophysical Research Communications*, *478*(3), 1268–1273.
- UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. Nucleic acids research, 47(D1), D506-D515.
- 26. Sondka, Z., Bamford, S., Cole, C. G., Ward, S. A., Dunham, I., & Forbes, S. A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer*, *18*(11), 696-705.
- 27. Landrum, M. J., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic acids research*, *46*(D1), D1062-D1067.
- 28. Uhlen, M., et al. (2010). Towards a knowledge-based human protein atlas. *Nature biotechnology*, *28*(12), 1248-1250.
- Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., ... & Leach, A.
 R. (2019). ChEMBL: towards direct deposition of bioassay data. *Nucleic acids research*, *47*(D1), D930-D940.
- Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., ... & Wilson,
 M. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research*, *46*(D1), D1074-D1082.
- 31. Rose, P. W., Prlić, A., Altunkaya, A., Bi, C., Bradley, A. R., Christie, C. H., ... & Burley, S.K. (2016). The RCSB protein data bank: integrative view of protein, gene and 3D structural

information. Nucleic acids research, gkw1000.

- 32. Hacker, S. M., Backus, K. M., Lazear, M. R., Forli, S., Correia, B. E., & Cravatt, B. F. (2017). Global profiling of lysine reactivity and ligandability in the human proteome. *Nature chemistry*, *9*(12), 1181-1190.
- Abbasov, M. E., Kavanagh, M. E., Ichu, T. A., Lazear, M. R., Tao, Y., Crowley, V. M., ... & Cravatt, B. F. (2021). A proteome-wide atlas of lysine-reactive chemistry. *Nature chemistry*, *13*(11), 1081-1092.
- 34. Braschi, B., Denny, P., Gray, K., Jones, T., Seal, R., Tweedie, S., ... & Bruford, E. (2019). Genenames. org: the HGNC and VGNC resources in 2019. *Nucleic acids research*, 47(D1), D786-D792.
- Gene Ontology Consortium. (2019). The gene ontology resource: 20 years and still GOing strong. *Nucleic acids research*, 47(D1), D330-D338.
- Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., ... & D'Eustachio, P. (2018). The reactome pathway knowledgebase. *Nucleic acids research*, 46(D1), D649-D655.
- 37. Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., ... & von Mering, C. (2021). The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*, *49*(D1), D605-D612.
- 38. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013;128(14)
- 39. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gundersen GW, Ma'ayan A. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research. 2016; gkw377.*

- 40. Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S., & Hopkins, A. L. (2012). Quantifying the chemical beauty of drugs. *Nature chemistry*, *4*(2), 90-98.
- 41. Benet, L. Z., Hosey, C. M., Ursu, O., & Oprea, T. I. (2016). BDDCS, the Rule of 5 and druggability. *Advanced drug delivery reviews*, *101*, 89-98.
- 42. Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. (2012). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, *64*, 4-17.
- 43. Ghose, A. K., Viswanadhan, V. N., & Wendoloski, J. J. (1999). A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *Journal of combinatorial chemistry*, 1(1), 55-68.
- 44. Congreve, M., Carr, R., Murray, C., & Jhoti, H. (2003). A'rule of three'for fragment-based lead discovery?. *Drug discovery today*, *8*(19), 876-877.
- 45. Landrum, G. (2013). Rdkit documentation. Release, 1(1-79), 4.
- Senkane, K., Vinogradova, E. V., Suciu, R. M., Crowley, V. M., Zaro, B. W., Bradshaw, J. M., ... & Cravatt, B. F. (2019). The Proteome-Wide Potential for Reversible Covalency at Cysteine. *Angewandte Chemie*, *131*(33), 11507-11511.
- 47. Krishnan, S., Miller, R. M., Tian, B., Mullins, R. D., Jacobson, M. P., & Taunton, J. (2014). Design of reversible, cysteine-targeted Michael acceptors guided by kinetic and computational analysis. Journal of the American Chemical Society, 136(36), 12624-12630.
- Zambaldo, C., Vinogradova, E. V., Qi, X., Iaconelli, J., Suciu, R. M., Koh, M., ... & Bollong,
 M. J. (2020). 2-Sulfonylpyridines as tunable, cysteine-reactive electrophiles. *Journal of the American Chemical Society*, *142*(19), 8972-8979.
- 49. Serafimova, I. M., Pufall, M. A., Krishnan, S., Duda, K., Cohen, M. S., Maglathlin, R. L., ...& Taunton, J. (2012). Reversible targeting of noncatalytic cysteines with chemically tuned

electrophiles. Nature chemical biology, 8(5), 471-476.

- 50. Du, X., Guo, C., Hansell, E., Doyle, P. S., Caffrey, C. R., Holler, T. P., ... & Cohen, F. E. (2002). Synthesis and structure– activity relationship study of potent trypanocidal thio semicarbazone inhibitors of the trypanosomal cysteine protease cruzain. Journal of medicinal chemistry, 45(13), 2695-2707.
- Greenbaum, D. C., Mackey, Z., Hansell, E., Doyle, P., Gut, J., Caffrey, C. R., ... & Chibale,
 K. (2004). Synthesis and structure– activity relationships of parasiticidal thiosemicarbazone cysteine protease inhibitors against Plasmodium falciparum, Trypanosoma brucei, and Trypanosoma cruzi. Journal of medicinal chemistry, 47(12), 3212-3219.
- 52. Shenai, B. R., Lee, B. J., Alvarez-Hernandez, A., Chong, P. Y., Emal, C. D., Neitz, R. J., ... & Rosenthal, P. J. (2003). Structure-activity relationships for inhibition of cysteine protease activity and development of Plasmodium falciparum by peptidyl vinyl sulfones. Antimicrobial agents and chemotherapy, 47(1), 154-160.
- 53. Klüver, E., Schulz-Maronde, S., Scheid, S., Meyer, B., Forssmann, W. G., & Adermann,
 K. (2005). Structure- activity relation of human β-defensin 3: influence of disulfide bonds and cysteine substitution on antimicrobial activity and cytotoxicity. Biochemistry, 44(28), 9804-9816.
- 54. Grzonka, Z., Jankowska, E., Kasprzykowski, F., Kasprzykowska, R., Lankiewicz, L., Wiczk, W., ... & Grubb, A. (2001). Structural studies of cysteine proteases and their inhibitors. Acta Biochimica Polonica, 48(1), 1-20.
- Zanon, P. R., Yu, F., Musacchio, P., Lewald, L., Zollo, M., Krauskopf, K., ... & Hacker, S.
 M. (2021). Profiling the proteome-wide selectivity of diverse electrophiles.
- 56. Dana, J. M., Gutmanas, A., Tyagi, N., Qi, G., O'Donovan, C., Martin, M., & Velankar, S. (2019). SIFTS: updated Structure Integration with Function, Taxonomy and Sequences

resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic acids research*, *47*(D1), D482-D489.

- 57. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L.,
 ... & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic acids research*, *49*(D1), D412-D419.
- 58. Julio, A. R., & Backus, K. M. (2021). New approaches to target RNA binding proteins. *Current opinion in chemical biology*, *62*, 13-23.
- 59. de la Cruz, J., Kressler, D., & Linder, P. (1999). Unwinding RNA in Saccharomyces cerevisiae: DEAD-box proteins and related families. *Trends in biochemical sciences*, 24(5), 192-198.
- 60. Aubourg, S., Kreis, M., & Lecharny, A. (1999). The DEAD box RNA helicase family in Arabidopsis thaliana. *Nucleic acids research*, *27*(2), 628-636.
- Patmore, D. M., Jassim, A., Nathan, E., Gilbertson, R. J., Tahan, D., Hoffmann, N., ... & Gilbertson, R. J. (2020). DDX3X suppresses the susceptibility of hindbrain lineages to medulloblastoma. *Developmental cell*, *54*(4), 455-470.
- 62. Andrisani, O., Liu, Q., Kehn, P., Leitner, W. W., Moon, K., Vazquez-Maldonado, N., ... & Gale, M. (2022). Biological functions of DEAD/DEAH-box RNA helicases in health and disease.
- Mi, H., Muruganujan, A., Ebert, D., Huang, X., & Thomas, P. D. (2019). PANTHER version
 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis
 tools. *Nucleic acids research*, *47*(D1), D419-D426.
- 64. Fesik, S. W. (2005). Promoting apoptosis as a strategy for cancer drug discovery. *Nature Reviews Cancer*, *5*(11), 876-885.
- 65. Aguilar, A., Lu, J., Liu, L., Du, D., Bernard, D., McEachern, D., ... & Wang, S. (2017). 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. *Journal of medicinal chemistry*, *60*(7), 2819-2839.

- 66. Giancotti, F. G., & Ruoslahti, E. (1999). Integrin signaling. *science*, 285(5430), 1028-1033.
- Cooper, J., & Giancotti, F. G. (2019). Integrin signaling in cancer: mechanotransduction, stemness, epithelial plasticity, and therapeutic resistance. *Cancer cell*, *35*(3), 347-367.Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, *33*(suppl_1), D514-D517.
- 68. Schrijver, I., Liu, W., Brenn, T., Furthmayr, H., & Francke, U. (1999). Cysteine substitutions in epidermal growth factor–like domains of fibrillin-1: distinct effects on biochemical and clinical phenotypes. *The American Journal of Human Genetics*, 65(4), 1007-1020.
- Russell, D. W., Brown, M. S., & Goldstein, J. L. (1989). Different combinations of cysteinerich repeats mediate binding of low density lipoprotein receptor to two different proteins. *Journal of Biological Chemistry*, 264(36), 21682-21688.
- 70. Daly, N. L., Scanlon, M. J., Djordjevic, J. T., Kroon, P. A., & Smith, R. (1995). Threedimensional structure of a cysteine-rich repeat from the low-density lipoprotein receptor. *Proceedings of the National Academy of Sciences*, *92*(14), 6334-6338.
- 71. Esser, V., Limbird, L. E., Brown, M. S., Goldstein, J. L., & Russell, D. W. (1988). Mutational analysis of the ligand binding domain of the low density lipoprotein receptor. *Journal of Biological Chemistry*, 263(26), 13282-13290.
- 72. Lanman, B. A., Allen, J. R., Allen, J. G., Amegadzie, A. K., Ashton, K. S., Booker, S. K., ... & Cee, V. J. (2019). Discovery of a covalent inhibitor of KRASG12C (AMG 510) for the treatment of solid tumors.

- Janes, M. R., Zhang, J., Li, L. S., Hansen, R., Peters, U., Guo, X., ... & Liu, Y. (2018).
 Targeting KRAS mutant cancers with a covalent G12C-specific inhibitor. *Cell*, *172*(3), 578-589.
- 74. Patricelli, M. P., Janes, M. R., Li, L. S., Hansen, R., Peters, U., Kessler, L. V., ... & Liu, Y. (2016). Selective Inhibition of Oncogenic KRAS Output with Small Molecules Targeting the Inactive StateTargeting Inactive KRASG12C Suppresses Oncogenic Signaling. *Cancer discovery*, *6*(3), 316-329.
- 75. Ostrem, J. M., Peters, U., Sos, M. L., Wells, J. A., & Shokat, K. M. (2013). K-Ras (G12C) inhibitors allosterically control GTP affinity and effector interactions. *Nature*, *503*(7477), 548-551.
- Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., ... & Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote) omics data. *Nature methods*, *13*(9), 731-740.
- 77. Perez-Riverol, Y., Bai, J., Bandla, C., García-Seisdedos, D., Hewapathirana, S., Kamatchinathan, S., ... & Vizcaíno, J. A. (2022). The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic acids research*, 50(D1), D543-D552.



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**.



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

View a Disease Related Protein

View a Detected Protein

(**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 publically available at https://backuslab.shinyapps.io/cysdb/.



b

5 CysDB Cysteines		0 Pathoge	nic Missense	e Variants		0 Cancer (Census Genes	
Protein Schematic with Labeled \	/ariants 👔							
	Amino	• 100 2 acid number	•••	200	•	 identified ligandabi hyperreaa in allele C range: 11 in allele C range: 12 in allele C range: 14 in dbSNP range: 84 in dbSNP range: 20 	e ttive STO1*B; decreased 5-155 STO1*C; no effect r 0-140 rrs11509436 -86 rrs11509438 8-206	i protein stability on protein stability; dbSNP
Identified Ligandable Rea	active							
Enrichment Options		Enric	nment of Ider	ntified Protein	ns Results			
Select an Identified Dataset or weerapana_cravatt	CysDB ID	Show	entries Term • Ov	rerlap ♦ P.value	Adjusted.P.v	ralue 🛊 Odds.Ratio 🕯	Combined.Score	Search: Genes
Select Enrichment Libra	ry	1 tra (C	anslation 50/3 60:0006412)	214 0	0	5.764	250.16	RPL30;RPLP1;RPLP0;
Submit		2 de (C	granulation 70/4 iO:0043312)	481 0	0	3.243	106.357	CDA;CYFIP1;GPI;PYG
		3 in re (0	eutrophil tivation volved in 70/- mune 50002283)	485 0	0	3.211	103.945	CDA;CYFIP1;GPI;PYG
		Showing	1 to 3 of 3,781 ent	ries		Previous	1 2 3 4	5 1,261
Individual Compound Physiochem	ical Properties	& Protein Ta	argets 👔					
ACRYL_24	Q	Necular Weight	_	log P		Top 25 Cysteines	s with the Highest	Ligandablility Ratios
Compound Structure	<u>4</u> 24	7.12	X	1.77	5 011 4 5		•	
	\$	oms	\$	Heavy Atoms 18	Competition Ra		••••	• • •
pline a	Ht 3	ond Acceptors		Hbond Donors 0	2.5	P04183 P04183 P01483 P01591 075607 043396 043396 043395 015235	Q7Z6Z7 Q00610 P83731 P227707 P13796 P13796 P10599	Q9Y4H4 Q9UHR5 Q99439 Q99439 Q99439 Q99439 Q99439 Q99439 Q994585 Q994585
							0000000	

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. The function tab provides general information on the POI, including subcellular locations, GO/KEGG terms and protein-protein interaction maps. In addition, centered on the function tab is a 'site map,' indicating which cysteines have been identified, liganded or hyperreactive by chemoproteomics. By clicking on the activity tab, one can assess the potential druggability of their POI through small-molecule binding annotations and heatmaps for quantitative chemoproteomic measures of hyper-reactivity and ligandability. For a comprehensive view of the structural environment surrounding the chemoproteomic detected cysteines, publicly available 3D crystals structures are displayed in the structure tab. Users can choose which structure is shown, add customized labels. (B) The disease-relevance of a POI can be explored through the mutation page. Proximity of chemoproteomic detected cysteines, annotated smallmolecule 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, orange for ligandable and pink 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 gueries 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. (E) Bar graph of the overall number of unique cysteines targeted by acrylamides or chloroacetamides or chloroacetamide. (F) Percentage of acrylamide and chloroacetamide compounds with a ratio >= 4 for protein carbonyl reductase (CBR1, UniProt: P16512) and protein glutathione s-transferase omega-1 (GSTO1, UniProt: 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) Distributions of ligandable (green) and hyperreactive (light blue) cysteines annotated as cysteine-specific binding sites (top) or cysteine-specific active sites (bottom). The total number of cysteines in UniProt annotated as binding or active sites are shown in gray. (C) Percentage of CysDB Lig (top) and hyperreactive (bottom) cysteine containing proteins with an associated PDB structure. (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 proteins associated with cancer by the Cancer Gene Census (CGC), CysDB ligandable proteins and 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_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 proteins associated with ClinVar variants, 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.