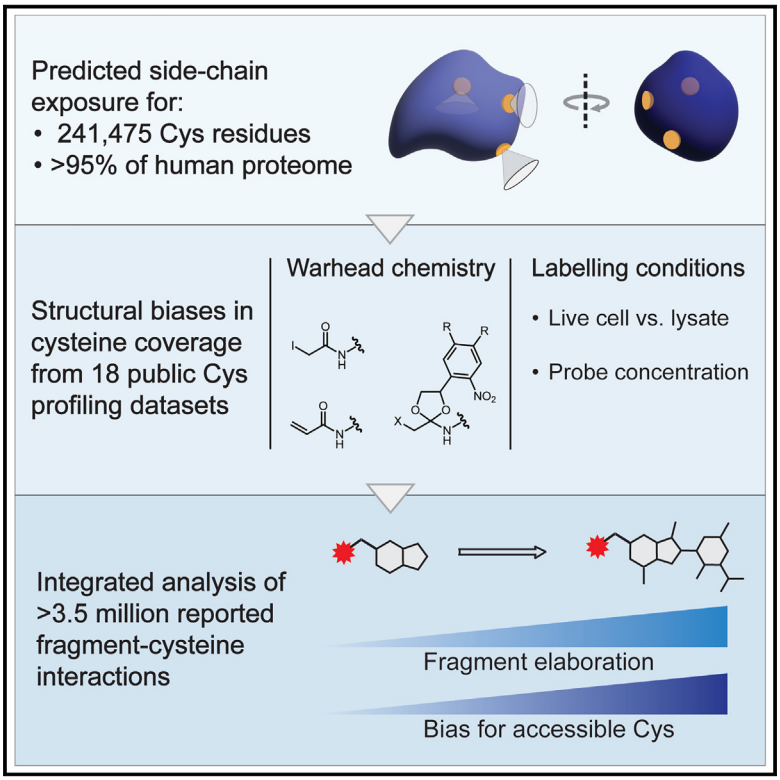
Resource

Cell Chemical Biology

Proteome-wide structural analysis identifies warhead- and coverage-specific biases in cysteine-focused chemoproteomics

Graphical abstract



Authors

Matthew E.H. White, Jesús Gil, Edward W. Tate

Correspondence

e.tate@imperial.ac.uk

In brief

White et al. perform a meta-analysis of public cysteine reactivity profiling studies integrated with AlphaFold2-predicted structures, identifying structural biases in cysteine profiling workflows and demonstrating inhibitor elaboration drives selectivity against buried cysteines. The reported results and analytical framework can guide development of future cysteine profiling approaches and covalent inhibitor design.

Highlights

- Amino acid side-chain accessibility (pPSE) for >95% of human proteome predicted
- Predicted structures capture pPSE distributions for functional and ligandable Cys
- Cysteine profiling workflows are biased by warhead and labeling conditions
- Fragment elaboration to drug-like compounds drives bias toward accessible Cys

White et al., 2023, Cell Chemical Biology *30*, 828–838
July 20, 2023
https://doi.org/10.1016/j.chembiol.2023.06.021

CellPress

CellPress
OPEN ACCESS

Cell Chemical Biology

Resource

Proteome-wide structural analysis identifies warhead- and coverage-specific biases in cysteine-focused chemoproteomics

Matthew E.H. White,[1,2] Jesús Gil,[2,3] and Edward W. Tate[1,4,5,*]
[1]Department of Chemistry, Molecular Sciences Research Hub, Imperial College London, London W12 0BZ, UK
[2]MRC London Institute of Medical Sciences (LMS), London W12 0NN, UK
[3]Institute of Clinical Sciences (ICS), Faculty of Medicine, Imperial College London, London W12 0NN, UK
[4]The Francis Crick Institute, London NW1 1AT, UK
[5]Lead contact
*Correspondence: e.tate@imperial.ac.uk
https://doi.org/10.1016/j.chembiol.2023.06.021

SUMMARY

Covalent drug discovery has undergone a resurgence over the past two decades and reactive cysteine profiling has emerged in parallel as a platform for ligand discovery through on- and off-target profiling; however, the scope of this approach has not been fully explored at the whole-proteome level. We combined AlphaFold2-predicted side-chain accessibilities for >95% of the human proteome with a meta-analysis of eighteen public cysteine profiling datasets, totaling 44,187 unique cysteine residues, revealing accessibility biases in sampled cysteines primarily dictated by warhead chemistry. Analysis of >3.5 million cysteine-fragment interactions further showed that hit elaboration and optimization drives increased bias against buried cysteine residues. Based on these data, we suggest that current profiling approaches cover a small proportion of potential ligandable cysteine residues and propose future directions for increasing coverage, focusing on high-priority residues and depth. All analysis and produced resources are freely available and extendable to other reactive amino acids.

INTRODUCTION

Covalent drug discovery has re-emerged over the past two decades as a powerful modality for difficult-to-drug and conventionally "intractable" targets. Covalent inhibition of protein targets takes advantage of the inherent reactivity of specific amino acid side chains, primarily cysteine, but with a continually expanding scope encompassing lysine, threonine, histidine, and electrophilic N-terminal modifications.[1–5] Covalent binding by a targeted covalent inhibitor (TCI) presents a number of advantages over non-covalent binding, including extended rather than equilibrium-limited residence time at a target site, potentially increased tractability of shallow binding pockets or intrinsically disordered regions, and selectivity driven by disease-associated amino acid mutations or post-translational modification (PTM) (Figure 1A).[6,7] These benefits are exemplified by development and Food and Drug Administration (FDA) approval of covalent inhibitors for high-priority cancer targets, including KRAS [G12C],[7,8] EGFR,[9] and BTK.[10] Even covalent ligands for amino acids outside of enzyme active sites may offer valuable starting points as allosteric inhibitors or for covalent bifunctional molecules which recruit effector proteins to neosubstrates (e.g. covalent proteolysis or deubiquitinase targeting chimeras, PROTACs[11,12] or DUBTACs[13]). However, systematic discovery of novel and developable covalent ligands remains a significant bottleneck due to the requirement for balanced reactivity and sufficient selectivity toward a single target amino acid against all other accessible amino acids displaying similar chemistry.

Technology platforms which permit quantitative profiling of covalent protein modifications have become powerful tools for TCI discovery and development.[6] Competition-based chemoproteomic methods such as isotopic tandem orthogonal proteolysis—activity-based protein profiling (isoTOP-ABPP), initially developed by Cravatt and co-workers,[1] have seen widespread application for highly parallel and versatile analysis of potential amino acid reactivity and ligandability. Such analyses initially focused on cysteine-reactive TCIs in cell lysates (*in vitro*) using a cysteine-reactive iodoacetamide warhead to enrich, identify, and quantify peptides bearing a cysteine existing at least partially in a reactive, and therefore potentially ligandable, state (Figure 1B, top).[1,2] Subsequent technical developments, including novel cleavable reagents, multiplexing strategies, chemical enrichment, and mass spectrometric (MS) acquisition approaches, have together markedly improved both depth and throughput, as reviewed comprehensively elsewhere.[14] Weerapana and co-workers further developed this concept for live-cell (*in situ*) labeling, using photo-uncaging of protected α-haloketones in live cells to permit labeling concentrations of up to 200 μM without significant cell

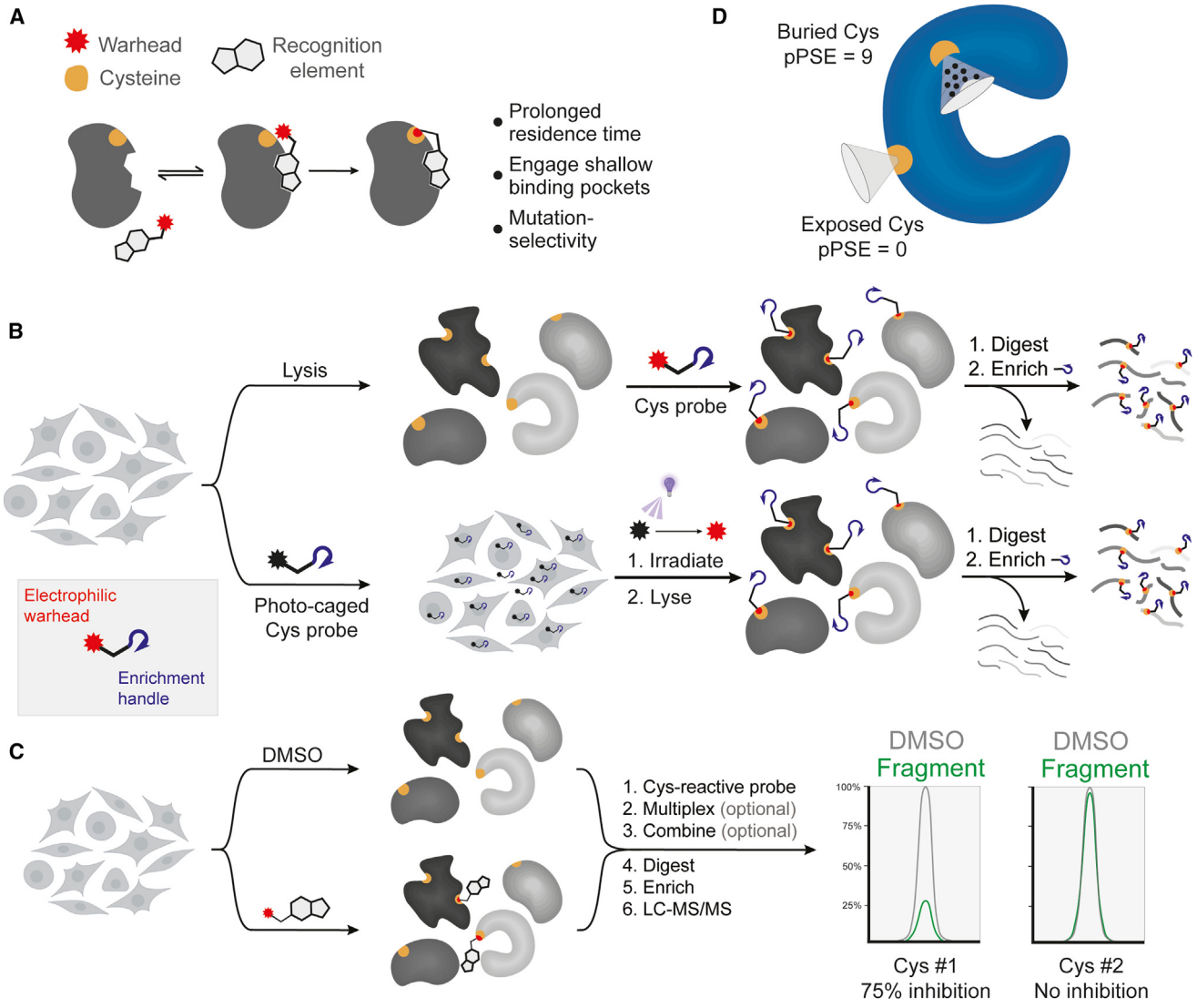


Figure 1. Overview of cysteine profiling approaches and AlphaFold2-based accessibility analysis applied in this study

(A) Representative mode of action for covalent binding by a targeted covalent inhibitor.

(B) Schematic showing cysteine profiling approaches applying peptide-level cysteine enrichment, either by labeling in cell lysates (top) or in live cells (bottom).

(C) Overview of competitive fragment screening by mass spectrometry-based cysteine profiling, with schematic chromatograms of fragment-competed and non-competed cysteines shown.

(D) Schematic for accessibility analysis on AlphaFold2-predicted structures as described by Bludau et al.[32].

death (Figure 1B, bottom).[15,16] Reactive amino acid profiling platforms have been applied extensively to covalent ligand discovery, primarily by quantifying cysteine reactivity in lysates treated with libraries of electrophilic fragments compared to DMSO-treated controls (Figure 1C). In these screens, loss of signal for a cysteine-containing peptide in a fragment-treated sample is interpreted as evidence for covalent fragment reactivity at that residue, with the magnitude of signal loss proportional to the occupancy of this interaction. In recent years, increasingly large-scale cysteine profiling (CP) experiments have generated a substantial body of publicly available data and subsequent resources allowing interpretation of chemoproteomic data in the context of experimental structures, clinical variants, and protein activity.[1,2,15–30]

The most comprehensive CP studies consistently profile 10,000 or more reactive cysteines in parallel. However, even considering a single canonical isoform per protein (as defined by the UniProt database), there are 261,260 cysteines in the human proteome and the structural distribution of reactive residues has yet to be systematically analyzed. Similarly, the extent to which CP approaches sample potentially ligand-accessible residues remains undefined but has significant implications for the efficiency of TCI discovery platforms and assessment of proteome-wide selectivity. Here, we present a meta-analysis of 18 published reactive cysteine datasets (Table S1) in the context of proteome-wide residue accessibility predictions enabled by AlphaFold2 (Figure 1D). We find significant variation in the accessibility distributions of profiled cysteines across

published studies, uncovering warhead-specific effects and clear disparities in residue targeting between cysteine-reactive enrichment probes, electrophilic fragment libraries, and optimized TCIs. We expect that these results, which we have collated in a publicly accessible and searchable structural database (https://tatelab.shinyapps.io/alpaca-db/), will help inform the applicability of different CP approaches at each stage of the probe or drug discovery pipeline, including direct screening by MS-based ABPP, target identification, and off-target profiling. We further suggest directions for future development of CP platforms to enhance and accelerate discovery of developable covalent ligands.

Amino acid side chains and post-translational modifications have distinct accessibility profiles

To predict the accessibility of amino acid side chains, we used a recently reported approach to integrate PTM proteomic datasets with AlphaFold2-predicted protein structural information across almost all human proteins.[31,32] In this approach, solvent accessibility calculations and predictions of folded/intrinsically disordered regions were integrated with phosphorylation, ubiquitination, and O-glycosylation datasets, and applied to uncover a number of PTM-specific structural distribution patterns. Of particular interest to the present study, the incorporation of AlphaFold2 prediction error into part-sphere exposure[33,34] calculations from predicted structures provides a proteome-wide metric of side-chain solvent accessibility at single-residue resolution. We calculated "prediction-aware part-sphere exposure" (pPSE) for each residue in 19,453 proteins from the human UniProt sequence database. The pPSE value of a given amino acid reflects the number of proximal α-carbons counted in a conical volume projecting 12 Å along the Cα-Cβ vector (or pseudo-vector in the case of glycine), with an internal angle of 70°. A high pPSE value represents a crowded environment around the side chain and therefore a more buried and less accessible residue, whereas a low pPSE value indicates high accessibility (Figures 1D and 2A).

We assessed the overall validity of this model across a range of relevant benchmarks. First, accessibility distributions show intuitive trends across intrinsic physicochemical and structural properties of individual amino acids. For example, charged amino acids such as Asp, Glu, Lys, and Arg show enrichment for highly exposed residues, whereas neutral and hydrophobic residues such as Val, Leu, Tyr, and Phe show a converse enrichment for buried residues (Figure 2B). Both Gly and Pro show a very strong enrichment for fully exposed (pPSE = 0) but not partially exposed residues (1 < pPSE <5), consistent with their known enrichment in short loop regions.[35] Focusing on subsets of annotated cysteine PTMs reveals distinct distributions for cysteine (Cys) annotated to engage in disulfide bonds, where interchain linkages are more exposed compared to more buried intrachain disulfides, consistent with accessibility to their respective disulfide partner (Figure 2C). Conversely, lipid-modified cysteine residues showed enrichment for highly exposed residues, with 63% of 1,084 *S*-acylated cysteines (from SwissPalm database) found at pPSE $\leq$5 and 100% of 219 known prenylated cysteines (from UniProt annotation[36]) at pPSE $\leq$3 (Figure 2C). pPSE distributions with high exposure are consistent with the requirement for enzymatic lipidation at cysteine, and mediation of subsequent membrane interactions at the protein surface.[37] Active site cysteine-annotated residues[36] showed greatest enrichment at 5 < pPSE <8, indicating the predicted "depth" of annotated active sites strikes an expected balance between substrate accessibility and solvent exclusion. Similar distributions are seen for both metal-binding and redox-active Cys.

Such wide coverage of structural models offers the opportunity to analyze amino acids proximal to Cys residues annotated with specific functions across thousands of cysteines. We calculated the pairwise enrichment based on Euclidean distance to each amino acid side chain around subsets of Cys residues, a three-dimensional analog of primary sequence enrichment analysis.[38] We observe intuitive enrichments proximal to both metal-binding Cys and active site Cys (Figure 2D) representative of the known role of both Cys and His in protein-metal interactions and, for example, Cys-His-Asp catalytic triads in active sites. In the subset of cysteines found to be hyperreactive toward iodoacetamide electrophiles, proximal Cys/His residues are enriched and both Asp and Glu are depleted (Figure S1A). These findings align with similar analysis performed with primary sequence enrichment of hyperreactive Cys, finding enrichment of proximal Cys, Trp, and Phe with varying distance in primary sequence.[39]

Together, these analyses indicate that physicochemically and biologically relevant trends for amino acid solvent accessibility can be extracted from computationally predicted structures alongside conservative filtering for prediction quality. However, we note that these predictions remain subject to the limitations of AlphaFold2 itself, particularly the absence of bound water molecules which may be directly involved at binding sites and in enzyme mechanisms, and caution should be exercised when drawing conclusions for any single residue in isolation in the absence of experimental validation.

Warhead chemistry and labeling environment generate distinct accessibility profiles

Depth of coverage of potentially ligandable cysteines remains important for CP workflows, as current acquisition approaches are limited to sampling a maximum of ~35,000 unique cysteines in a single experiment.[23] Although this coverage represents a tremendous improvement over first-generation technology, it encompasses around 20% of all MS-detectable cysteine-containing tryptic peptides (204,707, from Yan et al.[23]) and just 13% of all cysteines in the proteome.[23] Clearly, the subset of therapeutically relevant cysteines is significantly smaller, as a residue must be accessible to a small molecule, present at least partially in the nucleophilic thiol/thiolate form, and in the case of TCIs should result in a phenotypic change upon drug binding in a relevant physiological context. As such, enrichment of the most ligandable cysteines would achieve meaningful coverage of those residues most likely to prove fruitful as therapeutic targets, although there is currently no clear consensus on the size or character of this subset.

In CP experiments, preferential enrichment for surface residues has been proposed to occur by labeling with activity-based probes in minimally denatured lysates. For example, Backus et al. performed a direct comparison of cysteines quantified when labeling a "native lysate" with iodoacetamide-alkyne compared to those denatured by heating. Comparison of pPSE

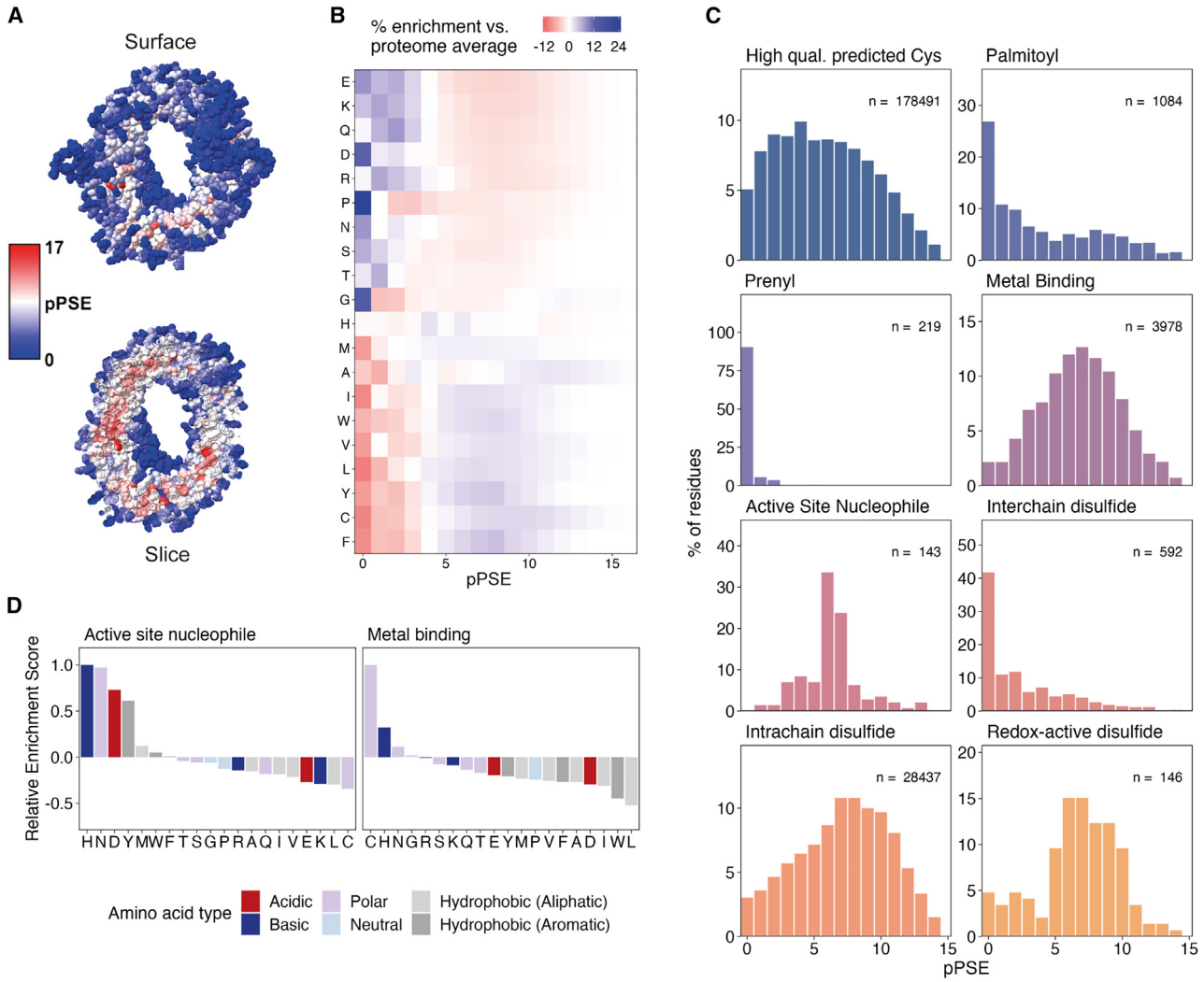


Figure 2. Amino acid- and PTM-specific structural environments from AlphaFold2-predicted structures

(A) AlphaFold2-predicted structure of XPO1 with atoms colored by accessibility (pPSE); bottom panel depicts the same structure sliced through to visualize buried residues and their pPSE values.

(B) Proteome-wide amino acid pPSE distributions normalized to whole-proteome average for all residues reflect side-chain physicochemical properties. Amino acids are ordered by percentage exposed (pPSE <5) residues.

(C) pPSE distributions of cysteine residues annotated with specific PTMs or functions in UniProt or SwissPalm.

(D) 3D (through-space) proximity enrichment of amino acids relative to cysteines with functional annotations reflects local environments conducive to metal binding or nucleophilic activity.

See also Figure S1.

distributions between these two strongly contrasted conditions demonstrates that residues with at least 2-fold higher enrichment in heat-treated samples ($R_{Native/Heat}$ < 0.5) also show increased pPSE, consistent with increased access to more buried cysteines (Figure 3A). Conversely, cysteines at least 2-fold enriched in the non-denatured lysate ($R_{Native/Heat}$ > 2) are enriched in exposed (pPSE <6) cysteines. A similar experiment reported by Li et al.[40] with detergent-based enrichment showed similar results (Figure 3B), confirming that minimally denaturing lysis preserves exposed cysteine labeling preferentially and that denaturation increases the accessibility of previously buried residues. Although the effect is less prominent, labeling with low (100 μM) and high (2 mM) concentrations of iodoacetamide-alkyne in lysates sampled differing accessibility distributions (Figures 3C and S2A, data from Yan et al.[23]) showing a shift toward sampling more buried residues at higher reagent concentrations.

We next sought to understand how the subset of cysteines sampled by a range of published CP protocols varied by accessibility. We compiled 18 published datasets, profiling a total of 44,187 unique cysteine residues and >3.5 million fragment-cysteine interactions across a range of warhead chemistries, enrichment strategies, and MS acquisition approaches (Table S1). We found that among iodoacetamide-based reagents, the overall subset of cysteine residues sampled is remarkably similar to the background distribution of cysteine

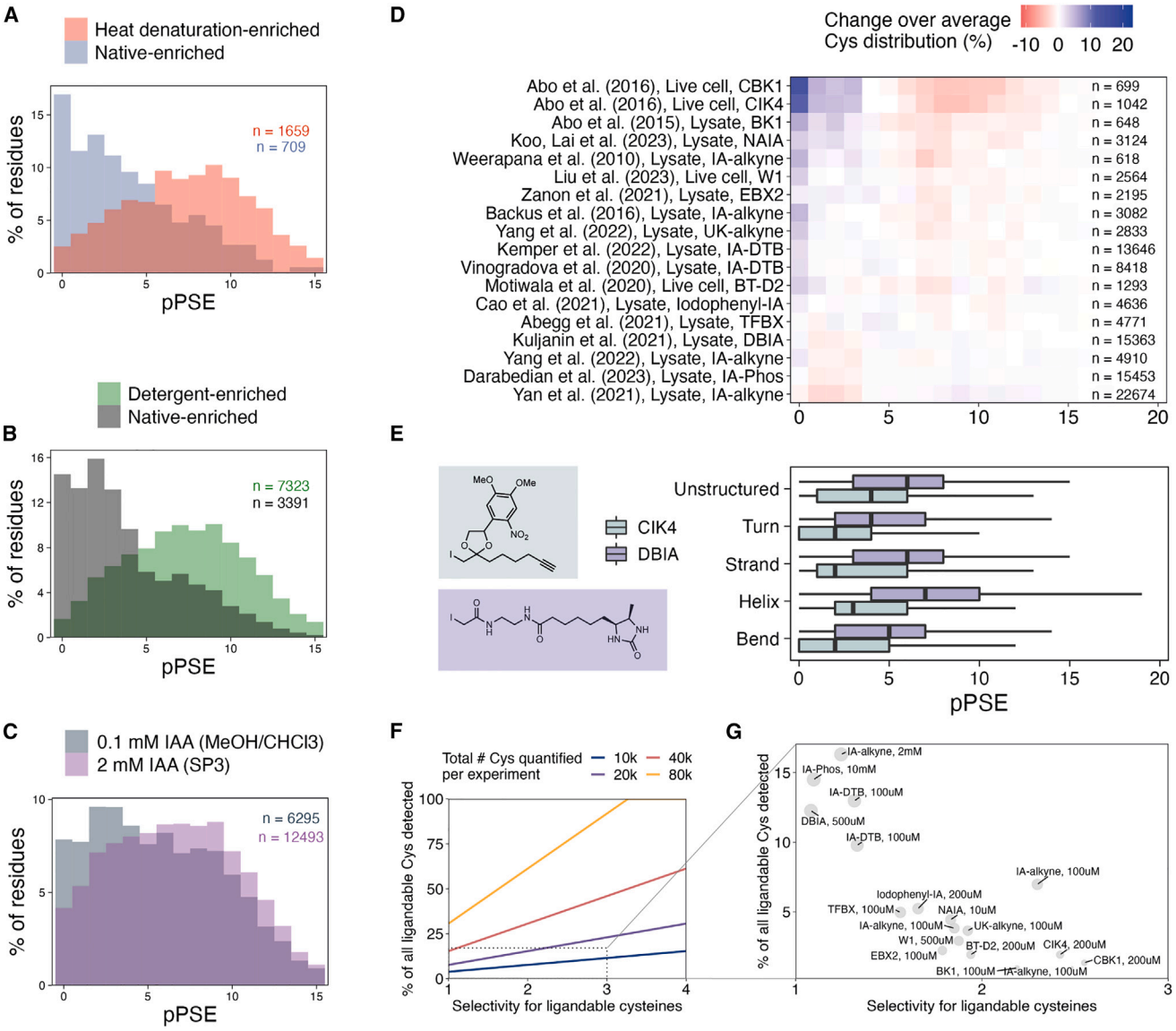


Figure 3. Cysteine profiling workflows sample different predicted accessibility distributions influenced by warhead chemistry and labeling conditions

(A) Effect of heat denaturation on accessibility of profiled cysteines. Native-enriched residues ($R_{Native/Heat} > 2$, blue) and denaturation-enriched residues ($R_{Native/Heat} < 0.5$, red) show distinct accessibility distributions; experimental data: Backus et al.[2].

(B) Effect of detergent denaturation on accessibility of profiled cysteines. Native-enriched residues ($R_{Native/Detergent} > 2$, gray) and denaturation-enriched residues ($R_{Native/Detergent} < 0.5$, green) show distinct accessibility distributions; experimental data: Li et al.[40].

(C) Effect of iodoacetamide-alkyne (IAA) labeling concentration on accessibility of profiled cysteines. All cysteines detected in experiments at either 0.1 or 2 mM are shown in gray and purple, respectively; experimental data: Yan et al.[23].

(D) Accessibility distributions for 18 reported CP datasets compared to the whole-proteome average distributions, ordered by bias toward more accessible Cys; colors denote percentage change of each pPSE over the whole proteome average.

(E) Comparison of iodoacetamide- and caged-iodoketone warhead across secondary structure groups as annotated by Bludau et al.[32]; experimental data: Kuljanin et al.[17] (desthiobiotin-iodoacetamide, DBIA) and Abo et al.[16] (caged-iodoketone, CIK4).

(F) Predicted coverage of potentially ligandable cysteines (y axis) at given overall cysteine coverage (lines), versus selectivity of labeling toward ligandable cysteines (x axis). Selectivity represents the fraction of liganded Cys identified in each study compared to the fraction of liganded Cys identified in combined fragment screening datasets (16.2%); the total number of ligandable Cys is estimated by extending the fraction of ligandable Cys (16.2%) to all 261,260 Cys in the human proteome.

(G) Inset from F, showing the coverage and selectivity values for 18 published CP datasets. For each point, the labeling probe and concentration are annotated; size of each point reflects the number of cysteines found in each experiment.

See also Figure S2.

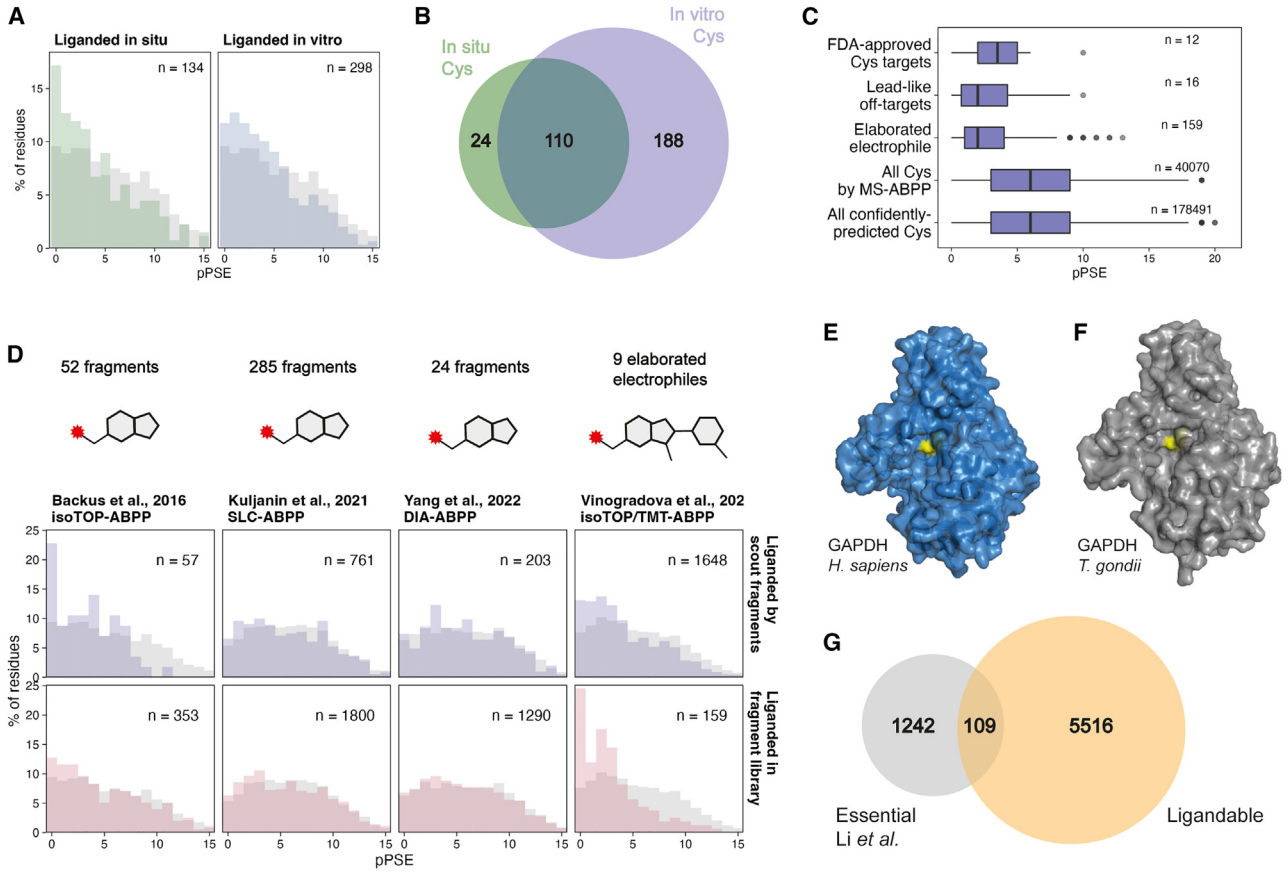


Figure 4. Covalent inhibitors, from small fragments to elaborated drug-like molecules, predominantly target exposed cysteines

(A) Covalent fragment-targeted cysteines are enriched for accessible cysteines over non-liganded residues (light gray), particularly when drug treatment is performed in live cells (*in situ*, left panel); experimental data from Backus et al.[2].

(B) Venn diagram showing the overlap between cysteines targeted ($R_{DMSO/Fragment} > 4$) by *in situ* and *in vitro* fragment treatments.

(C) Cysteine targets of FDA-approved TCIs, off-targets for lead-like compounds, and immunomodulatory elaborated electrophiles (Vinogradova et al.[18]) show bias toward accessible cysteine residues, compared to the distributions sampled by all described MS-ABPP datasets and the average of all cysteine residues across the proteome. Boxplots display 25th, 50th (median), and 75th percentile, whiskers display upper/lower limits of data; outliers are plotted as points.

(D) Comparison of accessibility distributions for four fragment screening datasets.[2,17–19] Panels show the subset of cysteine residues liganded by two promiscuous scout electrophiles (dark blue, top panels) and covalent fragment/inhibitor libraries (red, bottom panels). Background distributions of non-liganded Cys are plotted for each study in gray. Inset shows number of unique cysteine residues plotted in each histogram.

(E) AlphaFold2-predicted structure for *H. sapiens* GAPDH, with active site Cys152 highlighted in yellow.

(F) *T. gondii* GAPDH2 with Cys897 (corresponding to the conserved *H. sapiens* Cys152) highlighted in yellow; additional domains have been removed from visualization for clarity.

(G) Overlap of Cys defined as essential and in this study as ligandable (studies as referenced in panel D); experimental data from Li et al.[32].

See also Figure S3.

accessibility (Figure 3D). Notably, both N-acryloylindole[29] and α-bromoketone[15] (BK1) warheads showed modest enrichment of more exposed residues. The clearest enrichment is observed with photocaged α-haloketones (CIK4, CBK1), which permeate live cells without significant labeling and are then uncaged by prompt irradiation at 365 nm (as shown in Figure 1B, lower workflow).[15] Although lower in overall cysteine coverage (1,497 unique cysteines in total across both live-cell datasets), live-cell labeling with photocaged warheads shows significant bias for exposed residues across all annotated protein secondary structures (Figure 3E) and therefore is not driven merely by a preference for highly accessible, unstructured regions. The additional enrichment of accessible cysteines resulting from uncaging and live-cell labeling relative to BK1 labeling in lysates may be attributable to the native labeling cellular environment. These observations suggest that even mild, detergent-free lysis presents a measurable departure from native cellular conditions, for example by oxidation of highly reactive residues[41] or partial mechanical or thermal denaturation (e.g. by sonication). Furthermore, the timescales of uncaging and labeling likely differ in live cells compared to lysates, alongside probe quenching in redox-buffered live cells. Notably, studies with the highest enrichment for exposed residues, including with photocaged ligands, also tend to have a lower number of total identified Cys and the most commonly identified Cys between studies is more accessible (Figure S2B). This could be due to an inherent bias in detectability of exposed residues or confounding factors between different studies which specifically affect peptide

detection (variation in cysteine modification chemistry, fractionation strategies, LC-MS platform, or data analysis). We suggest that the chemical and biological origins of this differential labeling warrant further detailed investigation.

Taken together, these data indicate that probe chemistry and labeling conditions significantly affect the subset of cysteines sampled in a given experiment and should be critically evaluated based on the experimental design. For applications such as target identification from phenotypic covalent fragment screening or off-target profiling of developed TCIs, comprehensive (>50%) coverage of ligandable cysteines will be required to achieve a reasonable success rate using CP workflows alone. A number of computational studies have defined ligandable cysteines from analysis of cysteine orientation, adjacent residues, and binding pocket characterization;[42–44] however, these approaches have thus far relied on experimental structures. Defining ligandability instead from empirical fragment screening datasets,[2,17–19] we observed that 16.2% of profiled Cys (6,474/39,948) have at least one significant ($R_{Fragment/DMSO} \geq 4$) liganding event, relatively consistent with previous estimates.[2,17] Extending this proportion to all Cys in the proteome, a reasonable approximation given the aggregate of CP datasets to date closely matches the global Cys accessibility distribution (Figure 3D), yields an estimate of 42,350 ligandable Cys in the proteome. We therefore calculated the predicted percentage coverage of ligandable Cys as a function of both labeling selectivity and total number of Cys detected in a given experiment (Figure 3F). The current state of the art allows for up to 35,000 Cys in a single quantitative comparison;[21] however, given the observed minimal enrichment for ligandable Cys (selectivity ~1), this represents coverage of only 14.5% of all potentially ligandable Cys. To further investigate the extent to which diverse labeling conditions and warhead chemistries affect selectivity for ligandable Cys residues, we calculated the fraction of identified Cys in each study which are ligandable vs. non-liganded. We observed a clear correlation between enrichment for accessibility and selectivity toward ligandable Cys (Figure S2C), reaching a maximum selectivity for ligandable Cys of 2.5-fold with photocaged warheads (Figure 3G).

Consistent with recent advances in high-throughput proteome profiling and advances in instrumentation,[45–48] peptide detection and quantification improvements in the coming years will no doubt improve the number of cysteines profiled in a given MS-ABPP experiment. Combined with technical advances in sample preparation, acquisition, and analysis to alleviate known limitations of bottom-up approaches (e.g. multiple protease strategies, PTM-aware database searching), improvements in reproducible cysteine depth are to be expected; however, our results highlight the need to also optimize CP toward higher selectivity for potentially ligandable residues. Almost 40% coverage of the ligandable cysteinome is theoretically possible with the best reported coverage and selectivity in existing studies, although these factors may prove challenging to combine, as noted previously.

Reactive fragments and drug-like covalent inhibitors primarily target accessible residues

Finally, we performed a targeted re-analysis of combined fragment screening datasets with our accessibility analysis to provide a proteome-wide window on potentially ligandable cysteines. In the first report of covalent fragment screening by isoTOP-ABPP, a comparison of fragment treatments in lysates (*in vitro*) or live cells (*in situ*) was performed. Inspecting the accessibility distributions of the liganded Cys in each condition showed that the liganded cysteines in both treatments prioritize more accessible residue distributions when compared to all detected cysteines (Figure 4A), but with a greater bias in the live-cell treatment, suggesting that these highly accessible cysteines represent the most ligandable covalent targets in a native cellular environment. Although substantially more liganded cysteines were identified *in vitro* compared to *in situ* drug treatment (298 vs. 134), 82% of *in-situ*-identified sites (110 out of 134) were identified in both conditions (Figure 4B). Based on these findings, we compiled cysteine targets of FDA-approved TCIs,[49] off-targets of lead-like/FDA-approved inhibitors,[17,50–53] and immunomodulatory elaborated electrophiles,[18] showing enrichment for exposed cysteine residues compared to the distribution of cysteines detected by MS-ABPP or the proteome average (Figure 4C).

We further sought to understand the structural context of cysteines liganded in large-scale competition fragment experiments. A specific set of small, promiscuous electrophilic compounds, also termed "scout fragments" by analogy to the low molecular weight fragments applied in conventional fragment-based ligand discovery, have been applied in diverse biological contexts to determine differential proteome reactivity, for example upon activation of T cells[18] or NRF2-knockdown in non-small-cell lung cancer cells.[54] We therefore integrated reactivity data from two commonly profiled promiscuous fragments (KB02, KB05, Figure S3A) across 4 generations of CP technologies.[2,17–19] Two datasets (Backus et al. and Vinogradova et al.) show that scout-reactive cysteine profiles are enriched in exposed residues, compared to no enrichment for non-liganded cysteines, although this enrichment is observed to a lesser extent in the datasets by Kuljanin et al. and Yang et al. (Figure 4D, upper panels), at least in part due to predominantly non-overlapping cysteines sampled in each approach (Figures S3B–S3D). A similar trend is also evident in respective larger fragment screening experiments, where Backus et al. and Vinogradova et al. show enrichment for more exposed residues; however, such enrichment is less apparent in the datasets by Kuljanin et al. and Yang et al. (Figure 4D, lower panels). Categorizing liganded cysteines by warhead (chloroacetamide vs. acrylamide) from each fragment library showed no clear accessibility trends across the three datasets (Figure S3E). We further calculated various physicochemical properties for screened fragments (cLogP, cLogS, H-bond donors, H-bond acceptors, molecular weight, absolute and relative polar surface area, and total surface area)[55] and found that highly promiscuous fragments (>50 liganded residues) lie predominantly in the upper quartile of cLogP and lower quartile of cLogS values, consistent with promiscuous non-specific interactions of lipophilic fragments with protein surfaces (Figure S3F). 3D proximity analysis also identified local amino acids enriched in highly liganded Cys, mainly showing enrichment of polar or charged amino acids and depletion of neutral, hydrophobic residues (Figure S3G); however, this correlation is also seen for accessibility in general and therefore cannot be disentangled at the whole-proteome level. Taken

altogether, our results suggest that a subset of exposed cysteines represent the target residues for covalent fragments and, to an even greater extent, for more developed TCIs.

While we have focused on using predicted structures to determine proteome-wide trends in CP workflows and fragment engagement, functionality of specific Cys residues is also an important criterion for TCI development since it underpins the potential to disrupt function through ligand binding. Two recent studies have approached this problem with CRISPR-based screening for fitness upon mutation of specific Cys residues, opening up the possibility to identify therapeutically actionable Cys targets.[40,56] Combining such essentiality screens with both chemoproteomic and structural information can provide additional insights into TCI tractability of a particular Cys. For example, GAPDH2 Cys798 is found to be fitness-conferring for *Toxoplasma gondii* viability and corresponds to *Homo sapiens* GAPDH Cys152 by primary sequence alignment and overlay of predicted structures (Figures 4E and 4F). Indeed, Cys798 and Cys152 are found as hyperreactive residues in both *T. gondii*[56] and *H. sapiens*[20] hyperreactivity profiling experiments and have been identified in both human and *Escherichia coli* fragment screening datasets[57] (where *E. coli* Cys151 also corresponds to the active site Cys). There is clear evidence of homology and shared function between these residues, representing both a functional and ligandable site with structural homology observable by predicted structures. Thus, although the intersection between "essential" and experimentally ligandable Cys is relatively low (Figure 4G), it represents potentially high-value targets in a given physiological or pathological setting which could be greatly expanded by improvements in CP methodology coupled to more extensive functional Cys screening approaches.

DISCUSSION

Activity-based protein profiling technologies have made an important contribution to the covalent drug discovery pipeline, with cysteine-focused competitive profiling in particular offering a method for proteome-wide dissection of cysteine reactivity and ligandability. Our meta-analysis of publicly available cysteine profiling datasets and proteome-wide accessibility predictions shows that while different warhead chemistries sample distinct accessibility distributions, there is marked consistency across datasets with similar warheads. Furthermore, more specialized warhead chemistries show remarkable selectivities for both accessible and ligandable cysteines. We also observe enrichment for accessible residues in liganded targets of promiscuous scout fragments, fragment libraries, and drug-like elaborated electrophiles. The early stages of covalent ligand optimization campaigns are typically undertaken without the benefit of structure-guided design,[18,50,52] and we suggest that proteome-wide structural accessibility analysis may be a useful complementary approach when paired with CP workflows at each step of ligand optimization to probe the bias between accessible structural environments during evolution of a given ligand series.

Taken together, our analyses encourage further development of covalent ligand discovery workflows to enhance and optimize accessible residue coverage. Current high-throughput CP workflows spread coverage broadly across the full distribution of cysteine accessibility and appear to substantially under-sample potentially ligandable cysteines, even considering recent improvements in mass spectrometry which put 40,000+ quantified cysteines within reach. However, as noted previously, current MS technology is already capable of approaching full coverage of accessible cysteines if profiling capacity is tightly focused on accessible residues, for example by further development of in-cell labeling workflows driven by photocaged warheads or through combination of a defined set of less promiscuous Cys reactive molecules which together provide superior coverage of ligandable Cys residues. This consideration is equally important for *de novo* target deconvolution of bioactive compounds with a covalent mode of action, where incomplete coverage of ligandable cysteines greatly reduces the likelihood of positive target identification. Furthermore, as binding is inferred by loss of signal, there is potential for false positives through compound-induced changes in proximal PTMs, redox state, or protein conformation (and therefore cysteine accessibility) or simply sample handling. In these cases, positive target enrichment through direct labeling, for example with a bioactive probe bearing a clickable tag, is likely to be a more feasible approach.

We also identify several limitations of accessibility analysis which might be usefully addressed in the future. These include the potential for false positive or negative identification of exposed residues due to current limitations in structure prediction, which we have sought to minimize by limiting our global analysis to residues with high-confidence prediction and further limiting ligandability analysis to structured regions. While we believe our approach to be as useful and robust as the underlying AlphaFold2 data at a whole-proteome scale, any single prediction in isolation must be coupled to experimental validation to be considered actionable. The value of this analytical approach is expected to continually improve with evolution of machine learning approaches to structure prediction which take account of confounding factors, such as the presence of protein complex interfaces, protonation state, bound water molecules and solvation, or PTMs which might dramatically alter accessibility. Furthermore, it is important to recognize that accessibility is not synonymous with reactivity, and it is likely that a certain percentage of accessible residues are not amenable to liganding with warhead chemistries currently employed in profiling workflows; analyses will therefore benefit from refinement in parallel with the disclosure of increasingly large and diverse profiling datasets which reflect broader residue-level reactivity. A tighter integration with proteomic data might be achieved in future using filters which account for the limitations of proteomic analysis, for example the limited capability of proteomics workflows to deal with very short, very long, or highly modified peptides, encompassing a significant proportion of cysteines which may be important ligand target sites.[26]

Finally, we have made all resources/analysis freely available and created a tool for interactively visualizing collated site-specific chemoproteomic data on AlphaFold2 structures. We note that the analytical pipeline presented here is straightforward to implement and should be equally applicable to any residue-specific profiling pipeline. For example, as new profiling datasets become available, it will be interesting to observe the evolution of reactive amino acid coverage across models and species enabled by the 200 million predicted protein structures in the AlphaFold2 database, and by ongoing developments in warhead chemistries targeting residues beyond cysteine.[4,58]

Limitations of the study

Our meta-analysis demonstrates the power of combining chemoproteomic datasets with information gleaned from predicted protein structures, thanks to comprehensive coverage of the proteome by AlphaFold2 compared to current experimental structure repositories. Conversely, the predictive nature of AlphaFold2 presents its own limitations in accurately reflecting the complexity and dynamic nature of protein structures. We aimed to minimize the effect of low-confidence predictions by applying stringent filtering on prediction quality (pLDDT), confining ligandability analysis to structured domains, and drawing conclusions based on distributions of predicted values, rather than any one individual prediction. This approach will significantly benefit from future refinements in protein structure prediction, including extending analysis of accessibility to protein complexes and inclusion of bound water molecules. We combined 18 distinct CP datasets, representing a wide range of experimental parameters which are not individually controlled, including labeling concentrations, lysis conditions, cell lines, enrichment chemistry, and MS data acquisition and analysis. We therefore elected to take processed chemoproteomic datasets from each individual study as-published and subjected them to identical downstream filtering criteria to derive distribution-level comparisons. A further disparity, also highlighted by Palafox et al.,[20] is introduced in the mapping of identified cysteines to genomic coordinates and disparities in the proteome databases used for MS/MS matching in individual studies. In cases where only peptide sequences are provided in source publications, we matched the sequences to the canonical UniProt human FASTA. We have kept where possible the originally reported residue annotation (UniProt identifier and sequence position) from each dataset; however, as discussed previously, mismatches are inevitable and imply a subsequent filtering step in our analysis.

SIGNIFICANCE

Covalent drug discovery is a promising therapeutic modality, leveraging irreversible target occupancy, selectivity based on side-chain chemistry, and the potential to exploit previously unligandable binding sites. Current covalent ligand discovery most prominently targets cysteine residues and is frequently supported at multiple stages by cysteine-focused competitive chemoproteomic methods, which attempt to profile on- and off-target residues directly from hit compounds. We combined 18 published cysteine profiling datasets with side-chain accessibility predictions from AlphaFold2 to survey the landscape of cysteine-reactive binding for 44,187 unique cysteine residues. We found accessibility biases in residues sampled by cysteine profiling experiments based on warhead character, *in vitro* vs. *in situ* labeling, and probe concentration. We further considered the residues targeted in ABPP-based fragment screening studies across >3.5 million fragment-cysteine interactions and found that optimization of covalent fragments toward elaborated, drug-like compounds enhances the selectivity of covalent binding toward accessible cysteines, and against buried cysteines. Based on these findings, we suggest considerations for future development of cysteine profiling approaches to improve coverage of high-priority residues for ligand discovery, incorporating optimization of labeling chemistries as well as alleviating current limitations in sample preparation, data acquisition, and downstream analysis. We anticipate that our findings will be immediately useful in conjunction with existing cysteine profiling data, and widely applicable to reactive amino acid analyses for residues beyond cysteine. We finally provide a publicly available tool for visualizing ligandable and accessible cysteines on AlphaFold2 structures in conjunction with combined cysteine profiling datasets at https://tatelab.shinyapps.io/alpaca-db/.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- ADDITIONAL RESOURCES
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - pPSE prediction from human AlphaFold2 structures
  - Cysteine PTM and functional annotation
  - Cysteine profiling data curation
  - pPSE distribution analysis
  - Selectivity calculations
  - Fragment screening data curation
  - Fragment property analysis
  - Structure visualization
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.chembiol.2023.06.021.

ACKNOWLEDGMENTS

Figure 1B/C was created with BioRender.com. We are grateful to members of the Tate group at Imperial College and the Francis Crick Institute for their advice and suggestions during the refinement of this manuscript. This work was supported by Cancer Research UK with support from the Engineering and Physical Sciences Research Council (Program Award to EWT, DRCNPG-Nov21\100001; EPSRC CDT in Chemical Biology and CRUK Convergence Science Centre PhD studentship award to MEHW, EP/S023518/1/CANTAC721\100021 – Cancer Research UK Imperial Centre – Non-Clinical Training Award). Core support from MRC (MC_U120085810) and a grant from CRUK (C15075/A28647) funded research in J.G.'s laboratory.

AUTHOR CONTRIBUTIONS

Conceptualization, M.E.H.W., J.G., and E.W.T.; Methodology, M.E.H.W., J.G., and E.W.T.; Data Curation, Formal Analysis, and Investigation, M.E.H.W.; Resources, J.G. and E.W.T.; Writing – Original Draft, M.E.H.W., J.G., and E.W.T.; Writing – Review & Editing, M.E.H.W., J.G., and E.W.T.; Supervision, J.G. and E.W.T.; Funding Acquisition, J.G. and E.W.T.

DECLARATION OF INTERESTS

E.W.T. is or has been employed as a consultant or scientific advisory board member for Myricx Pharma, Samsara Therapeutics, Roche, Novartis, and Fastbase; research in his group has been funded by Pfizer Ltd, Kura Oncology, Daiichi Sankyo, Oxstem, Exscientia, Myricx Pharma, AstraZeneca, Vertex Pharmaceuticals, GSK, and ADC Technologies. E.W.T. holds equity in Myricx Pharma, Exactmer, and Samsara Therapeutics, and is a named inventor on patents filed by Myricx Pharma, Exactmer, Imperial College London, and the Francis Crick Institute. J.G. has acted as a consultant for Unity Biotechnology, Geras Bio, Myricx Pharma, and Merck KGaA. Pfizer and Unity Biotechnology have funded research in J.G.'s lab. J.G. owns equity in Geras Bio. J.G. is a named inventor in MRC and Imperial College patents.

Received: November 10, 2022
Revised: March 20, 2023
Accepted: June 23, 2023
Published: July 20, 2023

REFERENCES

1. Weerapana, E., Wang, C., Simon, G.M., Richter, F., Khare, S., Dillon, M.B.D., Bachovchin, D.A., Mowen, K., Baker, D., and Cravatt, B.F. (2010). Quantitative reactivity profiling predicts functional cysteines in proteomes. Nature *468*, 790–795.
2. 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., et al. (2016). Proteome-wide covalent ligand discovery in native biological systems. Nature *534*, 570–574.
3. Hacker, S.M., Backus, K.M., Lazear, M.R., Forli, S., Correia, B.E., and Cravatt, B.F. (2017). Global profiling of lysine reactivity and ligandability in the human proteome. Nat. Chem. *9*, 1181–1190.
4. Zanon, P.R.A., Yu, F., Musacchio, P., Lewald, L., Zollo, M., Krauskopf, K., Mrdović, D., Raunft, P., Maher, T.E., Cigler, M., et al. (2021). Profiling the proteome-wide selectivity of diverse electrophiles. Preprint at ChemRxiv. https://doi.org/10.26434/chemrxiv-2021-w7rss-v2.
5. Wang, X., Lin, Z., Bustin, K.A., McKnight, N.R., Parsons, W.H., and Matthews, M.L. (2022). Discovery of Potent and Selective Inhibitors against Protein-Derived Electrophilic Cofactors. J. Am. Chem. Soc. *144*, 5377–5388.
6. De Vita, E. (2021). 10 years into the resurgence of covalent drugs. Future Med. Chem. *13*, 193–210.
7. Ostrem, J.M., Peters, U., Sos, M.L., Wells, J.A., and Shokat, K.M. (2013). K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions. Nature *503*, 548–551.
8. Lanman, B.A., Allen, J.R., Allen, J.G., Amegadzie, A.K., Ashton, K.S., Booker, S.K., Chen, J.J., Chen, N., Frohn, M.J., Goodman, G., et al. (2020). Discovery of a Covalent Inhibitor of KRASG12C (AMG 510) for the Treatment of Solid Tumors. J. Med. Chem. *63*, 52–65.
9. Hossam, M., Lasheen, D.S., and Abouzid, K.A.M. (2016). Covalent EGFR Inhibitors: Binding Mechanisms, Synthetic Approaches, and Clinical Profiles. Arch. Pharm. *349*, 573–593.
10. Honigberg, L.A., Smith, A.M., Sirisawad, M., Verner, E., Loury, D., Chang, B., Li, S., Pan, Z., Thamm, D.H., Miller, R.A., et al. (2010). The Bruton tyrosine kinase inhibitor PCI-32765 blocks B-cell activation and is efficacious in models of autoimmune disease and B-cell malignancy. Proc. Natl. Acad. Sci. USA *107*, 13075–13080.
11. Tao, Y., Remillard, D., Vinogradova, E.V., Yokoyama, M., Banchenko, S., Schwefel, D., Melillo, B., Schreiber, S.L., Zhang, X., and Cravatt, B.F. (2022). Targeted Protein Degradation by Electrophilic PROTACs that Stereoselectively and Site-Specifically Engage DCAF1. J. Am. Chem. Soc. *144*, 18688–18699. https://doi.org/10.1021/jacs.2c08964.
12. Henning, N.J., Manford, A.G., Spradlin, J.N., Brittain, S.M., Zhang, E., McKenna, J.M., Tallarico, J.A., Schirle, M., Rape, M., and Nomura, D.K. (2022). Discovery of a covalent FEM1B recruiter for targeted protein degradation applications. J. Am. Chem. Soc. *144*, 701–708.
13. Henning, N.J., Boike, L., Spradlin, J.N., Ward, C.C., Liu, G., Zhang, E., Belcher, B.P., Brittain, S.M., Hesse, M.J., Dovala, D., et al. (2022). Deubiquitinase-targeting chimeras for targeted protein stabilization. Nat. Chem. Biol. *18*, 412–421.
14. Maurais, A.J., and Weerapana, E. (2019). Reactive-cysteine profiling for drug discovery. Curr. Opin. Chem. Biol. *50*, 29–36.
15. Abo, M., and Weerapana, E. (2015). A caged electrophilic probe for global analysis of cysteine reactivity in living cells. J. Am. Chem. Soc. *137*, 7087–7090.
16. Abo, M., Bak, D.W., and Weerapana, E. (2017). Optimization of caged electrophiles for improved monitoring of cysteine reactivity in living cells. Chembiochem *18*, 81–84.
17. 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., et al. (2021). Reimagining high-throughput profiling of reactive cysteines for cell-based screening of large electrophile libraries. Nat. Biotechnol. *39*, 630–641.
18. 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., et al. (2020). An activity-guided map of electrophile-cysteine interactions in primary human T cells. Cell *182*, 1009–1026.e29.
19. Yang, F., Jia, G., Guo, J., Liu, Y., and Wang, C. (2022). Quantitative chemoproteomic profiling with data-independent acquisition-based mass spectrometry. J. Am. Chem. Soc. *144*, 901–911.
20. Palafox, M.F., Desai, H.S., Arboleda, V.A., and Backus, K.M. (2021). From chemoproteomic-detected amino acids to genomic coordinates: insights into precise multi-omic data integration. Mol. Syst. Biol. *17*, e9840.
21. Darabedian, N., Ji, W., Fan, M., Lin, S., Seo, H.-S., Vinogradova, E.V., Yaron, T.M., Mills, E.L., Xiao, H., Senkane, K., et al. (2023). Depletion of creatine phosphagen energetics with a covalent creatine kinase inhibitor. Nat. Chem. Biol. *19*, 815–824. https://doi.org/10.1038/s41589-023-01273-x.
22. Motiwala, H.F., Kuo, Y.-H., Stinger, B.L., Palfey, B.A., and Martin, B.R. (2020). Tunable heteroaromatic sulfones enhance in-cell cysteine profiling. J. Am. Chem. Soc. *142*, 1801–1810.
23. Yan, T., Desai, H.S., Boatner, L.M., Yen, S.L., Cao, J., Palafox, M.F., Jami-Alahmadi, Y., and Backus, K.M. (2021). SP3-FAIMS chemoproteomics for high-coverage profiling of the human cysteinome. Chembiochem *22*, 1841–1851.
24. Cao, J., Boatner, L.M., Desai, H.S., Burton, N.R., Armenta, E., Chan, N.J., Castellón, J.O., and Backus, K.M. (2021). Multiplexed CuAAC Suzuki-Miyaura labeling for tandem activity-based chemoproteomic profiling. Anal. Chem. *93*, 2610–2618.
25. Yang, F., Chen, N., Wang, F., Jia, G., and Wang, C. (2022). Comparative reactivity profiling of cysteine-specific probes by chemoproteomics. Curr. Res. Chem. Biol. *2*, 100024.
26. Kemper, E.K., Zhang, Y., Dix, M.M., and Cravatt, B.F. (2022). Global profiling of phosphorylation-dependent changes in cysteine reactivity. Nat. Methods *19*, 341–352.
27. Abegg, D., Tomanik, M., Qiu, N., Pechalrieu, D., Shuster, A., Commare, B., Togni, A., Herzon, S.B., and Adibekian, A. (2021). Chemoproteomic profiling by cysteine fluoroalkylation reveals Myrocin G as an inhibitor of the nonhomologous end joining DNA repair pathway. J. Am. Chem. Soc. *143*, 20332–20342.
28. Liu, Y., Liu, J., Zhang, X., Guo, C., Xing, X., Zhang, Z.-M., Ding, K., and Li, Z. (2023). Oxidant-induced bioconjugation for protein labeling in live cells. ACS Chem. Biol. *18*, 112–122.
29. Koo, T.-Y., Lai, H., Nomura, D.K., and Chung, C.Y.-S. (2023). N-Acryloylindole-alkyne (NAIA) enables imaging and profiling new ligandable cysteines and oxidized thiols by chemoproteomics. Nat. Commun. *14*, 3564.
30. Boatner, L.M., Palafox, M.F., Schweppe, D.K., and Backus, K.M. (2023). CysDB: a human cysteine database based on experimental quantitative chemoproteomics. Cell Chem. Biol. *30*, 683–698.e3.

31. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589.

32. Bludau, I., Willems, S., Zeng, W.-F., Strauss, M.T., Hansen, F.M., Tanzer, M.C., Karayel, O., Schulman, B.A., and Mann, M. (2022). The structural context of posttranslational modifications at a proteome-wide scale. PLoS Biol. *20*, e3001636.

33. Hamelryck, T. (2005). An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. Proteins *59*, 38–48.

34. Heffernan, R., Dehzangi, A., Lyons, J., Paliwal, K., Sharma, A., Wang, J., Sattar, A., Zhou, Y., and Yang, Y. (2016). Highly accurate sequence-based prediction of half-sphere exposures of amino acid residues in proteins. Bioinformatics *32*, 843–849.

35. Krieger, F., Möglich, A., and Kiefhaber, T. (2005). Effect of proline and glycine residues on dynamics and barriers of loop formation in polypeptide chains. J. Am. Chem. Soc. *127*, 3346–3352.

36. UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. *49*, D480–D489.

37. Jiang, H., Zhang, X., Chen, X., Aramsangtienchai, P., Tong, Z., and Lin, H. (2018). Protein Lipidation: occurrence, mechanisms, biological functions, and enabling technologies. Chem. Rev. *118*, 919–988.

38. O'Shea, J.P., Chou, M.F., Quader, S.A., Ryan, J.K., Church, G.M., and Schwartz, D. (2013). pLogo: a probabilistic approach to visualizing sequence motifs. Nat. Methods *10*, 1211–1212.

39. Wang, H., Chen, X., Li, C., Liu, Y., Yang, F., and Wang, C. (2018). Sequence-based prediction of cysteine reactivity using machine learning. Biochemistry *57*, 451–460.

40. Li, H., Remsberg, J.R., Won, S.J., Zhao, K.T., Huang, T.P., Lu, B., Simon, G.M., Liu, D.R., and Cravatt, B.F. (2022). Assigning functionality to cysteines by base editing of cancer dependency genes. Preprint at bioRxiv. https://doi.org/10.1101/2022.11.17.516964.

41. Yamamoto, T., Suzuki, T., Kobayashi, A., Wakabayashi, J., Maher, J., Motohashi, H., and Yamamoto, M. (2008). Physiological significance of reactive cysteine residues of Keap1 in determining Nrf2 activity. Mol. Cell Biol. *28*, 2758–2770.

42. Du, H., Jiang, D., Gao, J., Zhang, X., Jiang, L., Zeng, Y., Wu, Z., Shen, C., Xu, L., Cao, D., et al. (2022). Proteome-wide profiling of the covalent-druggable cysteines with a structure-based deep graph learning network. Research *2022*, 9873564.

43. Awoonor-Williams, E., and Rowley, C.N. (2018). How reactive are druggable cysteines in protein kinases? J. Chem. Inf. Model. *58*, 1935–1946.

44. Zhao, Z., Liu, Q., Bliven, S., Xie, L., and Bourne, P.E. (2017). Determining cysteines available for covalent inhibition across the human kinome. J. Med. Chem. *60*, 2879–2889.

45. Meier, F., Brunner, A.-D., Frank, M., Ha, A., Bludau, I., Voytik, E., Kaspar-Schoenefeld, S., Lubeck, M., Raether, O., Bache, N., et al. (2020). diaPASEF: parallel accumulation-serial fragmentation combined with data-independent acquisition. Nat. Methods *17*, 1229–1236.

46. Bache, N., Geyer, P.E., Bekker-Jensen, D.B., Hoerning, O., Falkenby, L., Treit, P.V., Doll, S., Paron, I., Müller, J.B., Meier, F., et al. (2018). A novel LC system embeds analytes in pre-formed gradients for rapid, ultra-robust proteomics. Mol. Cell. Proteomics *17*, 2284–2296.

47. Messner, C.B., Demichev, V., Bloomfield, N., Yu, J.S.L., White, M., Kreidl, M., Egger, A.-S., Freiwald, A., Ivosev, G., Wasim, F., et al. (2021). Ultra-fast proteomics with Scanning SWATH. Nat. Biotechnol. *39*, 846–854. https://doi.org/10.1038/s41587-021-00860-4.

48. Ishikawa, M., Konno, R., Nakajima, D., Gotoh, M., Fukasawa, K., Sato, H., Nakamura, R., Ohara, O., and Kawashima, Y. (2022). Optimization of Ultrafast Proteomics Using an LC-Quadrupole-Orbitrap Mass Spectrometer with Data-Independent Acquisition. J. Proteome Res. *21*, 2085–2093.

49. Du, H., Gao, J., Weng, G., Ding, J., Chai, X., Pang, J., Kang, Y., Li, D., Cao, D., and Hou, T. (2021). CovalentInDB: a comprehensive database facilitating the discovery of covalent inhibitors. Nucleic Acids Res. *49*, D1122–D1129.

50. Niessen, S., Dix, M.M., Barbas, S., Potter, Z.E., Lu, S., Brodsky, O., Planken, S., Behenna, D., Almaden, C., Gajiwala, K.S., et al. (2017). Proteome-wide Map of Targets of T790M-EGFR-Directed Covalent Inhibitors. Cell Chem. Biol. *24*, 1388–1400.e7.

51. Patricelli, M.P., Janes, M.R., Li, L.-S., Hansen, R., Peters, U., Kessler, L.V., Chen, Y., Kucharski, J.M., Feng, J., Ely, T., et al. (2016). Selective Inhibition of Oncogenic KRAS Output with Small Molecules Targeting the Inactive State. Cancer Discov. *6*, 316–329.

52. Kavanagh, M.E., Horning, B.D., Khattri, R., Roy, N., Lu, J.P., Whitby, L.R., Ye, E., Brannon, J.C., Parker, A., Chick, J.M., et al. (2022). Selective inhibitors of JAK1 targeting an isoform-restricted allosteric cysteine. Nat. Chem. Biol. *18*, 1388–1398. https://doi.org/10.1038/s41589-022-01098-0.

53. Wijeratne, A., Xiao, J., Reutter, C., Furness, K.W., Leon, R., Zia-Ebrahimi, M., Cavitt, R.N., Strelow, J.M., Van Horn, R.D., Peng, S.-B., et al. (2018). Chemical Proteomic Characterization of a Covalent KRASG12C Inhibitor. ACS Med. Chem. Lett. *9*, 557–562.

54. Bar-Peled, L., Kemper, E.K., Suciu, R.M., Vinogradova, E.V., Backus, K.M., Horning, B.D., Paul, T.A., Ichu, T.-A., Svensson, R.U., Olucha, J., et al. (2017). Chemical Proteomics Identifies Druggable Vulnerabilities in a Genetically Defined Cancer. Cell *171*, 696–709.e23.

55. Sander, T., Freyss, J., von Korff, M., and Rufener, C. (2015). DataWarrior: an open-source program for chemistry aware data visualization and analysis. J. Chem. Inf. Model. *55*, 460–473.

56. Benns, H.J., Storch, M., Falco, J.A., Fisher, F.R., Tamaki, F., Alves, E., Wincott, C.J., Milne, R., Wiedemar, N., Craven, G., et al. (2022). CRISPR-based oligo recombineering prioritizes apicomplexan cysteines for drug discovery. Nat. Microbiol. *7*, 1891–1905.

57. Zanon, P.R.A., Lewald, L., and Hacker, S.M. (2020). Isotopically Labeled Desthiobiotin Azide (isoDTB) Tags Enable Global Profiling of the Bacterial Cysteinome. Angew. Chem. Int. Ed. Engl. *59*, 2829–2836.

58. Gilbert, K.E., Vuorinen, A., Aatkar, A., Pogány, P., Pettinger, J., Grant, E.K., Kirkpatrick, J.M., Rittinger, K., House, D., Burley, G.A., et al. (2023). Profiling Sulfur(VI) Fluorides as Reactive Functionalities for Chemical Biology Tools and Expansion of the Ligandable Proteome. ACS Chem. Biol. *18*, 285–295.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
DataWarrior	https://openmolecules.org/datawarrior/	v5.5.0
ChimeraX	https://www.rbvi.ucsf.edu/chimerax/	v1.5
PyMol	https://pymol.org/2/	v2.5.4
StructureMap	Bludau et al.[32]	https://github.com/MannLabs/structuremap
R	https://www.r-project.org/	v4.1.3
RStudio	https://rstudio.com/	Version 2022.02.1 + 461
ChemDraw 21.0.0	Perkin Elmer Informatics	https://perkinelmerinformatics.com/products/research/chemdraw
Other		
Analysis and resources associated with accessibility analysis of public cysteine profiling data	This paper	https://github.com/TateLab/cys-accessibility
Resource website for accessibility and ligandability visualization	This paper	https://tatelab.shinyapps.io/alpaca-db/
StructureMap accessibility calculations	This paper	https://doi.org/10.5281/zenodo.7752842
Quantitative reactivity profiling predicts functional cysteines in proteomes	41586_2010_BFnature09472_MOESM203_ESM.pdf (Table S4)	Weerapana et al.[1] PMID: 21085121
A chemoproteomic platform to quantitatively map targets of lipid-derived electrophiles	41592_2014_BFnmeth2759_MOESM638_ESM.xlsx	Wang et al.[5,39] PMID: 24292485
A Caged Electrophilic Probe for Global Analysis of Cysteine Reactivity in Living Cells	ja5b04350_si_001.xlsx	Abo et al.[15,16] PMID: 26020833
Optimization of Caged Electrophiles for Improved Monitoring of Cysteine Reactivity in Living Cells	cbic201600524-sup-0001-misc_information.xlsx	Abo et al.[15,16] PMID: 27813293
Proteome-wide covalent ligand discovery in native biological systems	41586_2016_BFnature18002_MOESM54_ESM.xlsx	Backus et al.[2] PMID: 27309814
Selective Inhibition of Oncogenic KRAS Output with Small Molecules Targeting the Inactive State	21598290cd151105-sup-155462_2_supp_3294592_q0fwqr.xlsx	Patricelli et al.[51] PMID: 26739882
An activity-guided map of electrophile-cysteine interactions in primary human T cells	1-s2.0-S0092867420308230-mmc4.xlsx	Vinogradova et al.[18] PMID: 32730809
Tunable Heteroaromatic Sulfones Enhance in-Cell Cysteine Profiling	ja9b08831_si_003.xlsx	Motiwala et al.[22] PMID: 31881155
Profiling the proteome-wide selectivity of diverse electrophiles	zanon-et-al-supplementary-Table S3 xlsx zanon-et-al-supplementary-Table S4 xlsx	Zanon et al.[4,57] https://doi.org/10.26434/chemrxiv-2021-w7rss-v2
Multiplexed CuAAC Suzuki-Miyaura Labeling for Tandem Activity-Based Chemoproteomic Profiling	ac0c04726_si_002.xlsx	Cao et al.[24] PMID: 33470097
SP3-FAIMS Chemoproteomics for High-Coverage Profiling of the Human Cysteinome	cbic202000870-sup-0001-table_s5.xlsx table_s1.xlsx cbic202000870-sup-0001-table_s2.xlsx	Yan et al.[23] PMID: 33442901

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Reimagining high-throughput profiling of reactive cysteines for cell-based screening of large electrophile libraries	41587_2020_778_MOESM6_ESM.xlsx 41587_2020_778_MOESM8_ESM.xlsx 41587_2020_778_MOESM8_ESM.xlsx 41587_2020_778_MOESM10_ESM.xlsx	Kuljanin et al.[17] PMID: 33398154
Chemoproteomic Profiling by Cysteine Fluoroalkylation Reveals Myrocin G as an Inhibitor of the Non-homologous End-Joining DNA Repair Pathway	ja1c09724_si_001.xlsx	Abegg et al.[27] PMID: 34817176
Quantitative Chemoproteomic Profiling with Data-Independent Acquisition-Based Mass Spectrometry	ja1c11053_si_003.xlsx	Yang et al.[19,25] PMID: 34986311
Global profiling of phosphorylation-dependent changes in cysteine reactivity	41592_2022_1398_MOESM2_ESM.xlsx	Kemper et al.[26] PMID: 35228727
From chemoproteomic-detected amino acids to genomic coordinates: insights into precise multi-omic data integration	msb20209840-sup-0020-datasetev18.xlsx	Palafox et al.[20] PMID: 33599394
Comparative reactivity profiling of cysteine-specific probes by chemoproteomics	1-s2.0-S2666246922000064-mmc1.xlsx	Yang et al.[19,25] https://doi.org/10.1016/j.crchbi.2022.100024
N-Acryloylindole-alkyne (NAIA) enables profiling new ligandable hotspots in chemoproteomics experiments and imaging thiol oxidation	supplementary-data-2.xlsx	Lai et al.[29] https://doi.org/10.26434/chemrxiv-2022-khww5
Depletion of creatine phosphagen energetics with a covalent creatine kinase inhibitor	41589_2023_1273_MOESM3_ESM.xlsx	Darabedian et al.[21] PMID: 36823351
Oxidant-Induced Bioconjugation for Protein Labeling in Live Cells	cb2c00740_si_002.xlsx	Liu et al.[28] PMID: 36543757
Assigning functionality to cysteines by base editing of cancer dependency genes	media-2.xlsx media-4.xlsx	Li et al.[40] https://doi.org/10.1101/2022.11.17.516964

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Edward W. Tate (e.tate@imperial.ac.uk)

Materials availability

This study did not generate new unique reagents.

Data and code availability

This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table. All original code and data/resources to reproduce all analysis has been deposited at Zenodo and GitHub and is publicly available as of the date of publication. DOIs are listed in the key resources table. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ADDITIONAL RESOURCES

An interactive viewer of AlphaFold2 structures, including visualization tools for highlighting specific Cys residues and a database of compiled fragment screening data, as well as fragment structure visualization is freely available at: https://tatelab.shinyapps.io/alpaca-db/

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

All data are generated from the datasets provided in the key resources table.

METHOD DETAILS

pPSE prediction from human AlphaFold2 structures

Solvent accessibility prediction, intrinsically-disordered region prediction and secondary structure annotations were calculated using StructureMap as described in Bludau et al.[32] (https://github.com/MannLabs/structuremap). In brief, for each amino acid a conical volume is extended out from the β-carbon (α-carbon in the case of Gly) with an internal angle of 70° and distance of 12 Å. The number of proximal amino acid α-carbon atoms which sit within this cone (within the tolerance of the predicted aligned error) is then counted. This count is the pPSE value reported. Code to reproduce this analysis is publicly available at www.github.com/TateLab.

Cysteine PTM and functional annotation

Residue-specific annotations were downloaded from UniProt, including functions (metal binding, active site nucleophile) and PTMs (prenyl, disulfide) and matched to unique UniProt accession IDs and residue positions. Palmitoyl modifications were downloaded from SwissPalm (https://swisspalm.org/) No filtering for prediction quality was applied across the pPSE distributions. For all mappings, the UniProt canonical human FASTA (downloaded 26/10/2022) was used. The reference proteome and formatted databases used for these analyses are provided at: https://zenodo.org/record/7752842.

Cysteine profiling data curation

All datasets were downloaded as Supplemental Information files from their respective publications (Table S1, key resources table). For analysis of cysteines detected in each dataset, all unique cysteine residues were extracted and matched to AlphaFold2-predicted pPSE values by UniProt accession ID and sequence position. Any reverse/contaminant proteins and proteins with no predicted pPSE values were removed and cysteines were then filtered by the following criteria: ambiguous cysteine identifications (i.e. annotations with >1 possible cysteine) and low-confidence predictions (AlphaFold2 prediction quality <70) were removed. Additionally, for coverage (Figure 3) and fragment ligandability analysis (Figure 4) cysteines predicted to be in unstructured protein domains were removed.

pPSE distribution analysis

For all distributions, the number of residues at each pPSE value was normalized by the total number of cysteines per experiment/condition such that the sum of all pPSE fractions was 1. All confidently predicted cysteine residues (quality >70) were used for reference distributions and relative change (Figure 3D) was calculated by subtracting the reference distribution from each dataset (at respective pPSE values).

Selectivity calculations

The number of ligandable cysteines was determined by calculating the maximum R value per unique Cys residue across four reported fragment screening datasets. Cys with $R_{Max} \geq 4$ were annotated as liganded and the proportion of ligandable Cys (16.2%) was used in later calculation. For each CP dataset, the fraction of previously annotated ligandable Cys was calculated out of all fragment-screened Cys (excluding Cys not found in fragment screening studies). As there is significant difference in the cysteines samples by each warhead/dataset, the fraction of ligandable Cys was then extended to all Cys for each dataset (including those not found in fragment screening data) to estimate the number of ligandable Cys identified. Percentage coverage of ligandable Cys was then calculated for each dataset relative to the estimated number of ligandable Cys in the proteome (42,350). Selectivity was defined as the fraction of ligandable Cys relative to the fraction from all ligandable Cys (16.2%).

Fragment screening data curation

Fragment screening datasets were downloaded as Supplemental Information files from their respective publications (Table S1). Integration with AlphaFold2-predicted pPSE values and subsequent filtering was performed as described above, except for selectivity analysis where no filtering for quality or accessibility annotation was applied. Fragment-cysteine interactions with $R_{DMSO/Fragment} > 4$ were annotated as liganded.

Fragment property analysis

SMILES strings were manually collated from Supplemental Information files of each fragment screening dataset and imported into OSIRIS DataWarrior[55] (v 5.5.0, Actelion Ltd). Physicochemical properties were calculated using DataWarrior built-in functions and exported to R for visualization.

Structure visualization

AlphaFold2 structures were downloaded from the Alphafold Protein Structure Database (https://alphafold.ebi.ac.uk/). For Figure 2A, B-factors, representing prediction quality (pLDDT) as downloaded were replaced with pPSE values and visualized with UCSF ChimeraX (v1.5, https://www.rbvi.ucsf.edu/chimerax/). For Figures 4E and 4F, AlphaFold2 structures were visualized in PyMol (v2.5.4, Schrödinger, LLC).

QUANTIFICATION AND STATISTICAL ANALYSIS

No statistical tests were applied in our analysis. Sample sizes for distributions can be found inset in each figure.