CIAA: Integrated Proteomics and Structural Modeling for Understanding Cysteine Reactivity with Iodoacetamide Alkyne

Lisa M. Boatner[1,2], Jerome Eberhardt[3], Flowreen Shikwana[1,2], Matthew Holcomb[3], Peiyuan Lee[4], Kendall N. Houk[2], Stefano Forli[3] and Keriann M. Backus[1,2,5,6,7]

1. Biological Chemistry Department, David Geffen School of Medicine, UCLA, Los Angeles, CA, 90095, USA.
2. Department of Chemistry and Biochemistry, UCLA, Los Angeles, CA, 90095, USA.
3. Department of Integrative Structural and Computational Biology, Scripps Research Institute, La Jolla, CA, 92037, USA.
4. Department of Statistics and Data Science, UCLA, Los Angeles, CA, 90095, USA.
5. DOE Institute for Genomics and Proteomics, UCLA, Los Angeles, CA 90095, USA
6. Jonsson Comprehensive Cancer Center, UCLA, Los Angeles, CA 90095, USA
7. Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, UCLA, Los Angeles, CA 90095 USA

Correspondence: forli@scripps.edu or kbackus@mednet.ucla.edu

Abstract

Cysteine residues play key roles in protein structure and function and can serve as targets for chemical probes and even drugs. Chemoproteomic studies have revealed that heightened cysteine reactivity towards electrophilic probes, such as iodoacetamide alkyne (IAA), is indicative of likely residue functionality. However, while the cysteine coverage of chemoproteomic studies has increased substantially, these methods still only provide a partial assessment of proteome-wide cysteine reactivity, with cysteines from low abundance proteins and tough-to-detect peptides still largely refractory to chemoproteomic analysis. Here we integrate cysteine chemoproteomic reactivity datasets with structure-guided computational analysis to delineate key structural features of proteins that favor elevated cysteine reactivity towards IAA. We first generated and aggregated multiple descriptors of cysteine microenvironment, including amino acid content, solvent accessibility, residue proximity, secondary structure, and predicted pKa. We find that no single feature is sufficient to accurately predict reactivity. Therefore, we developed the CIAA (Cysteine reactivity towards IodoAcetamide Alkyne) method, which utilizes a Random Forest model to assess cysteine reactivity by incorporating descriptors that characterize the 3D structural properties of thiol microenvironments. We trained the CIAA model on existing and newly generated cysteine chemoproteomic reactivity data paired with high-resolution crystal structures from the Protein Data Bank (PDB), with cross validation against an external dataset. CIAA analysis reveals key features driving cysteine reactivity, such as backbone hydrogen bond donor atoms, and reveals still underserved needs in the area of computational predictions of cysteine reactivity, including challenges surrounding protein structure selection dataset curation. Thus our work provides a strong foundation for deploying artificial intelligence (AI) on cysteine chemoproteomic datasets.

Introduction

Cysteine residues are privileged sites in proteins, acting as redox sensors, catalytic nucleophiles, structural motifs, and even targets of chemical probes and FDA approved drugs.[1–5] Consequently, the identification of functional and potentially druggable cysteines is a central challenge of functional biology and drug development. The intrinsic reactivity of the cysteine thiol side chain towards electrophilic reagents has emerged as a key parameter that correlates with both functionality and druggability.[6] While the pKa of a thiol is around 8.5,[1] the pKa of a cysteine's thiol side chain can vary significantly depending on protein microenvironment (pKa 3.5 to 10), the reactivity of cysteines towards minimalized electrophilic molecules, such as iodoacetamide alkyne (IAA), is both time- and concentration-dependent.[7]

Measurements of cysteine reactivity have been generated proteome-wide, using the chemoproteomic method, isotopic Tandem Orthogonal Proteolysis-Activity-Based Protein Profiling (isoTOP-ABPP). For these analyses cysteine reactivity is assessed by quantifying the relative labeling with high (10x) versus low (1x) concentrations of IAA, using a proteomic readout. Highly reactive, or "hyper-reactive," cysteines are those that show a similar labeling with high and low IAA concentrations, ($\text{Ratio}_{[\text{high}]/[\text{low}]}$ =1), indicating saturation of labeling at the lower IAA concentration. High-reactivity has been found to be indicative of cysteine functionality, including involvement in catalytic activity and susceptibility to oxidative modifications.[8,9] Further illustrating the functional relevance of these measurements, our recent work revealed an enrichment for high predicted pathogenicity (high CADD score) for the codons of high reactivity cysteines.[10]

Despite the considerable value of these reactivity measurements, coverage remains a major challenge that has yet to be fully addressed. Reactivity measurements are currently only available for ~1.5% of all cysteines.[6,10,11] However, ~78% of cysteines should be theoretically detectable based on tryptic peptide length (>6 & <45 amino acids).[12] Reasons for this incomplete coverage include protein sequences that differ from reference sequences, genes with restricted expression, cysteines that are buried or in structural disulfides, and ionization properties of peptides.

Computational predictions of cysteine reactivity represent an exciting strategy to pinpoint functional residues, in a manner complementary to chemoproteomic analysis. [13–17]tructure-based programs like PROPKA[18] and H++[19] can predict pKa values with variable accuracy. Advances such as Cy-preds[20] and GB-CpHMD[21] incorporate both sequence and 3D structural data, but their application remains limited to a small set of protein structures and conformations.[22–24] Stepping beyond these smaller datasets, machine learning applied to chemoproteomics datasets has proven useful in identifying primary sequence motifs correlated with cysteine reactivity.[13–17] Whether the addition of 3D structural information can enhance the performance of such models remains to be seen. While not yet applied to reactivity analysis, the availability of *in silico* packages for covalent docking at cysteine residues[25–29] points towards as yet untapped opportunities for integrating reactivity measurements with protein structures to further guide discovery of reactive cysteine residues.

Here we establish the CIAA (Cysteine reactivity towards IodoAcetamide Alkyne) platform, which is tailored to guide the *in silico* discovery of high reactivity cysteines. To build CIAA we first generated a high coverage proteomic datasets of high reactivity cysteines that features 823 total high reactivity cysteines, identified in both newly generated and previously published datasets. We achieve >50% increase in total high reactivity cysteines when compared to prior datasets. We then subject a class-balanced set of high- and low reactivity cysteines to feature analysis, both in linear sequence and 3D protein space. While we find several features that are suggestive of cysteine reactivity, including most notably frequent proximity to histidine and proline residues, no single feature showed a strong correlation with cysteine high reactivity. Therefore, we developed a Random Forest model that was trained on 3D protein structures from the Protein Data Bank

(PDB). The model integrates curated chemoproteomic datasets with additional publicly available datasets, creating a robust framework for training. Validated with external datasets achieved an overall accuracy of 68%. Notable features identified by the model as correlated with cysteine reactivity include backbone hydrogen bond donor atoms, proximity to pockets and intermediate values of solvent accessibility. Taken together we expect that the CIAA platform will facilitate ongoing and future efforts towards high accuracy *in silico* discovery of functional and potentially druggable cysteine residues.

Results

Establishing a high coverage dataset of high reactivity cysteines. Our first step towards enhancing the *in silico* discovery of high reactivity cysteines, was to generate a high coverage dataset of known cysteines that exhibit a range of reactivities towards the pan-cysteine reactive probe iodoacetamide alkyne (IAA). We opted to pursue a hybrid strategy, both aggregating previously reported datasets[6,10] together with production of new in-house generated proteome-wide measures of cysteine reactivity. For both previously acquired and our newly generated datasets, relative intrinsic cysteine reactivity towards IAA was quantified by comparing labeling with either high (100 µM) or low (10 µM) concentration IAA, with saturation of labeling at lower probe concentration indicative of cysteine high reactivity.

For our reanalysis, we curated a set of cysteine high reactivity data that had previously been generated using the Isotopic Tandem Orthogonal Proteolysis-Activity-Based Protein Profiling (isoTOP-ABPP) chemoproteomic sample preparation method (**Figure 1A**).[6] Samples analyzed by isoTOP-ABPP were reprocessed for Weerapana et al. 2010 (n = 6)[6] and Palafox et al. 2021 (n = 5).[10] Reanalysis was conducted to ensure consistency in processing, address reproducibility, and confirm high-confidence identification of high reactivity cysteines across datasets. In total these prior datasets contained 489 total high reactivity cysteines, defined as $R_{[\text{high IAA}]/[\text{low IAA}]} = R_{100:10}$ values ≤ 2.3, with the remaining 8,115 total cysteines categorized as either medium (2.3 < $R_{10:1}$ values < 10), or low reactivity ($R_{10:1}$ values ≥ 10).

Given the comparatively modest size of the reanalysis dataset—the human proteome harbors ~260,000 cysteines by comparison[30]–we also generated additional in-house reactivity analysis (n = 13) for proteome derived from the HEK293T cell line. HEK293T cells are a commonly used workhorse cell line that has not to our knowledge been subjected to such reactivity analysis. These new datasets added 4204 cysteines that had not been observed in prior hyperreactivity studies, 640 of which were found to be hyperreactive (**Figure 1B** and **Data S1**). In aggregate across both the newly generated and reanalyzed data, the relative reactivity of 9,783 cysteines from 3,974 proteins were quantified. Of these, ~80% of residues (7,964) showed medium reactivity, with ~10% of cysteines exhibiting either high- or low-reactivity towards IAA (823 cysteines from 717 proteins and 996 cysteines from 803 proteins, respectively; **Data S1**).

Cysteine reactivity correlates with UniProtKB indications of functionality

As our newly generated data more than doubled the total number of high reactivity cysteines identified to-date (**Figure 1B**), we further benchmarked this new data to ensure that quality was maintained during this scale-up process. We observe a good overlap in cysteines identified (3,445 total shared) and a positive correlation between our new dataset and those previously reported (Pearson correlation coefficient 0.5, **Figure S1**). Consistent with prior reports of cell-line dependent differences in cysteine reactivity and ligandability,[6,31] we do note some likely cell-type specific differences in reactivity, for example cysteine 140 in Inosine-5'-monophosphate dehydrogenase 2 (IMPDH2). In addition to comparing ratio concordance between datasets, we also assessed whether previously reported properties of high reactivity cysteines were maintained in our new and larger dataset. Notably, and corroborating prior findings[6], we observe that cysteine high reactivity provides a good metric of likely functional significance, as indicated by the enrichment for residues in functional sites, including active sites, redox sensitive sites and disulfides, with the latter expected to be redox-active disulfides (**Figure 1C** and **Data S1**). Intriguingly, our UniProtKB analysis also revealed a notable correlation between low reactivity residues and metal binding sites, including zinc fingers (**Figure S2**). In total, 30 low reactivity cysteines were identified with UniProtKB annotations related to zinc binding or zinc finger regions, compared to 20 high reactivity cysteines. This analysis confirmed that our newly generated data did extend cysteine coverage while showing a similar properties distribution of previously reported datasets.

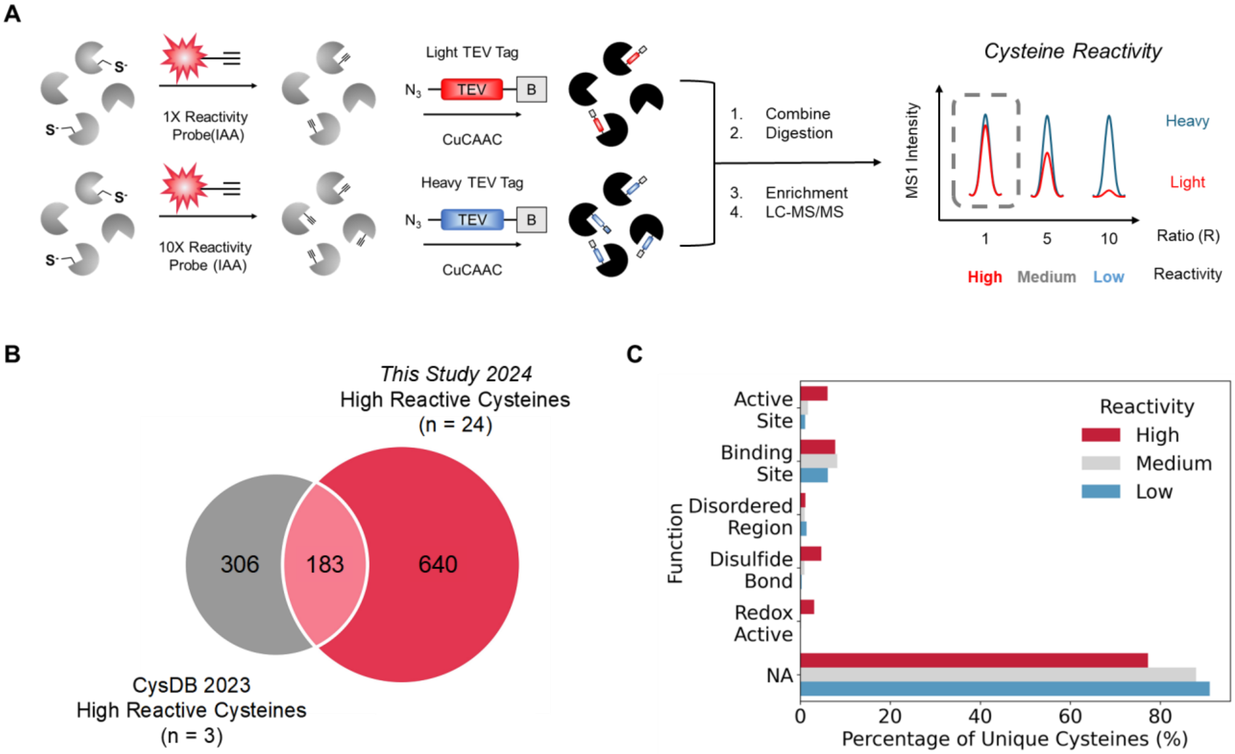


Figure 1. Establishing a dataset of high reactivity cysteines towards iodoacetamide alkyne (IAA). (A) Experimental workflow for isoTOP-ABPP. Cell lysates are treated with either high (100 µM) or low (10 µM) concentration of this IAA probe followed by click conjugation to isotopically differentiated tobacco etch virus (TEV)-cleavable biotinylated enrichment tags. After single-pot

solid phase-enhanced sample preparation (SP3) cleanup[12,32] and on-resin sequence-specific digestion, samples were enriched (streptavidin), eluted with TEV protease and the labeled peptides subjected to LC-MS/MS analysis followed by search with MSFragger,[33] using the FragPipe user interface and MS1-based quantification with IonQuant.[34] MS1 ratios correspond to $R_{heavy/light} = R_{[100\ \mu M]/[10\ \mu M]}$ with the following cutoffs for reactivity, high ($R_{100:10} \leq 2.3$), medium ($2.3 < R_{100:10} < 10$), and low ($R_{100:10} \geq 10$). (B) Comparison of the number of high reactivity cysteines identified in prior studies as reported in CysDB V1[30] for Weerapana et al. 2010[6], Palafox et al. 2021[10], and Vinogradova et al. 2020 versus high reactivity cysteines identified in newly generated datasets (n = 13). High-reactive cysteines were required to be identified in two replicates and had a $R_{100:10}$ standard deviation of <= 3 for further data analysis. (C) Comparison of UniProtKB functional annotations for high- vs low reactivity cysteines. See also **Figure S1**, **Figure S2**, and **Data S1**.

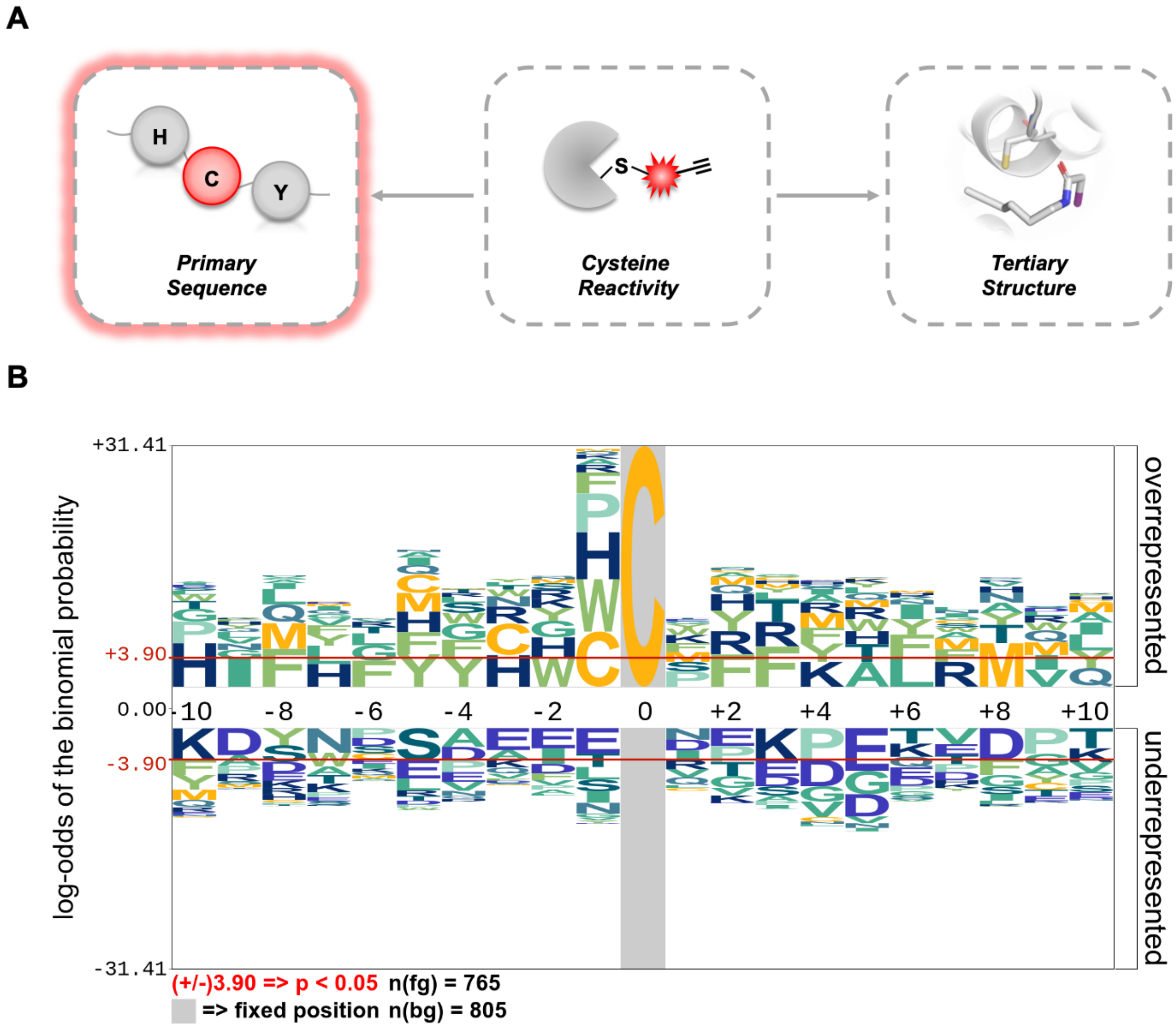


Figure 2. Amino acid contents of IAA-reactive cysteines using primary sequences. (A)

Overview of using primary sequences of IAA-labeled cysteines. (B) Sequence logo created using pLogo (http://plogo.uconn.edu).[35] Starting with 823 high reactivity and 996 low reactivity cysteines, sequences were aligned to meet pLogo input requirements, reducing the dataset to 765 high reactivity cysteines as the foreground and 805 low reactivity cysteines as the background. (A) shows the primary sequence motifs for these cysteines. The y-axis represents the log-odds binomial probability of an amino acid residue at a specific position, while the x-axis shows the position relative to a reactive cysteine fixed at position 0. The red horizontal line indicates the statistical significance threshold (p = 0.05) after applying the Bonferroni correction. See **Data S1**.

Primary sequence amino acid composition of high reactivity cysteines

Previous analysis of a focused set (n = 74) of high reactivity cysteines had revealed enrichment for tryptophan, histidine, proline, and cysteine residues in linear sequence proximity to high reactivity sites.[13,14] Therefore, to further assess how our dataset compares to this prior study and, particularly, to characterize whether these sequence-based enrichments hold true for our larger dataset, we next subjected our data to sequence motif analysis (**Figure 2A**). We generated a sequence logo using pLogo[35] to assess the frequencies of amino acids flanking high- and low reactivity cysteines, starting with 823 high reactivity cysteines and 996 low reactivity cysteines. After aligning the sequences to ensure they were of the same size and length, as required by the pLogo software, the dataset was reduced to 765 high reactivity cysteines as the foreground and 805 low reactivity cysteines as the background (**Figure 2B** and **Data S1**). This analysis revealed an increased occurrence of cysteines (C) near high reactivity cysteines at specific positions. At position -3, cysteines were slightly increased, with a log-odds of 4.1, consistent with a CXXC motif observed in the thioredoxin family.[36] A larger increase was found at position -1, with a log-odds of 7.2, indicating cysteines are most over-represented at this position. In addition to cysteines, histidine (H) and proline (P) were frequently found at position -1, while hydrophobic residues such as tryptophan (W), phenylalanine (F), and methionine (M) were identified within high reactivity cysteine neighborhoods. Acidic residues, including glutamate (E) and aspartate (D), were depleted, likely due to incompatible electrostatic interactions with cysteine thiolates. These trends are generally consistent with the aforementioned prior studies,[13,14] which indicates that sequence-based analysis likely can provide some indication of relative cysteine reactivity.

Defining a training set of reactive cysteines with 3D structural data available in the PDB

As one of the key overarching goals of our study is to define structural features that drive cysteine high reactivity, our next step was to step beyond linear sequence and to associate protein structural information with our identified cysteines (**Figure 3A**). Of our entire reactivity dataset, 66% (2636/3969) of the IAA-labeled proteins identified had experimentally determined protein structures deposited in the PDB (**Figure S3**). Similarly, 67% of proteins containing high reactivity cysteine proteins were structurally resolved (483/717) (**Figure 3B**). To check for potential biases in the representation of structures for the different protein families and for different cysteine reactivity classes, we analyzed their distribution in the available PDB structures. We found that the distribution of proteins with PDB structures closely resembles the distribution of proteins in the proteome with PDB structures and those experimentally labeled by IAA (**Figure S4**). We observed an enrichment of enzyme structures among proteins with PDB structures, while proteins without associated structures showed a higher prevalence of uncategorized proteins.

Many proteins still remain incompletely resolved and so some of our identified cysteines could be located in unresolved protein regions. Therefore, we next further filtered our dataset to ensure that all detected cysteines were structurally resolved. We matched the residue numbering and coordinates in the PDB files with UniProtKB amino acid numbering using custom scripts (see **Supplementary Computational Methods**). 345 out of 823 (42%) high reactivity cysteines and 322 out of 996 (33%) low reactivity cysteines were resolved in at least one corresponding crystal structure (**Figure 3C**).

To establish our curated training set, we opted to subject these structures to several additional pre-processing steps. Among these, we ensured that the IAA-reactive cysteine and its +/- 3 neighboring residues were fully resolved, with no missing density. This was a crucial step to achieve a comprehensive representation of the local microenvironment surrounding each cysteine. To exclude possible confounding effects of mutations or other protein modifications, we additionally excluded structures harboring these features from further analysis. Through these filtering steps, we also noted that nearly half of all proteins (241/505) had more than one associated structure in the PDB, with a small subset matching to >20 structures (**Figure S5A** and **Figure S5B**). To reduce the potential for data redundancy, we used the PISCES[37] server (accessed November 2023), which prioritized X-ray structures by selecting representatives based on structural quality and sequence diversity. This filtering reduced our set of structures from 22,821 to 1,179 PDBs, including 306 high reactivity and 297 low reactivity cysteines across 644 and 662 unique PDBs (**Figure 3C** and **Figure S5A**). Notably, 32 of these proteins contained both a high reactivity and a low reactivity cysteine. Importantly, and demonstrating that our filtering steps did not introduce significant bias to the datasets, the high reactivity and low reactivity protein sets exhibited similar distributions of experimental techniques, structural resolutions, and biological complexes (**Figure S6**).

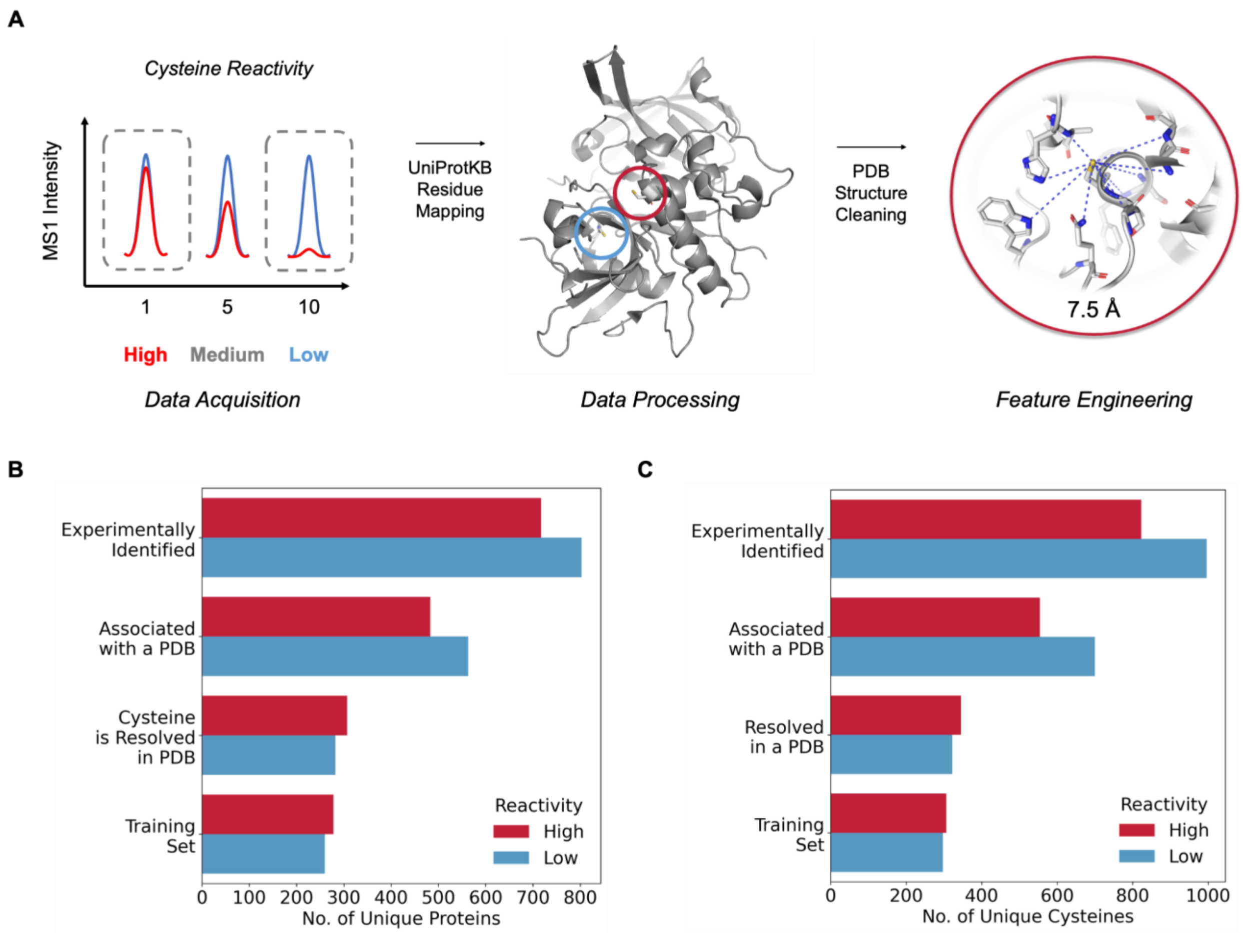


Figure 3. Defining a training set of reactive cysteines with 3D structural data available in the PDB. (A) Workflow for defining a training set of tertiary structures. (B) Number of experimentally identified proteins containing high- or low reactivity cysteines, number of experimentally identified unique high- or low reactivity cysteines associated with PDB structures, number of experimentally identified unique high- or low reactivity cysteines resolved in at least one associated PDB structure, and number of experimentally identified unique high- or low reactivity cysteines in the training set after a series of filtering steps. (C) Number of experimentally identified unique high- or low reactivity cysteines, number of experimentally identified unique high- or low reactivity cysteines associated with PDB structures, number of experimentally identified unique high- or low reactivity cysteines resolved in at least one associated PDB structure, and number of experimentally identified unique high- or low reactivity cysteines in the training set after a series of filtering steps. See **Figure S3-S6**, and **Data S2**.

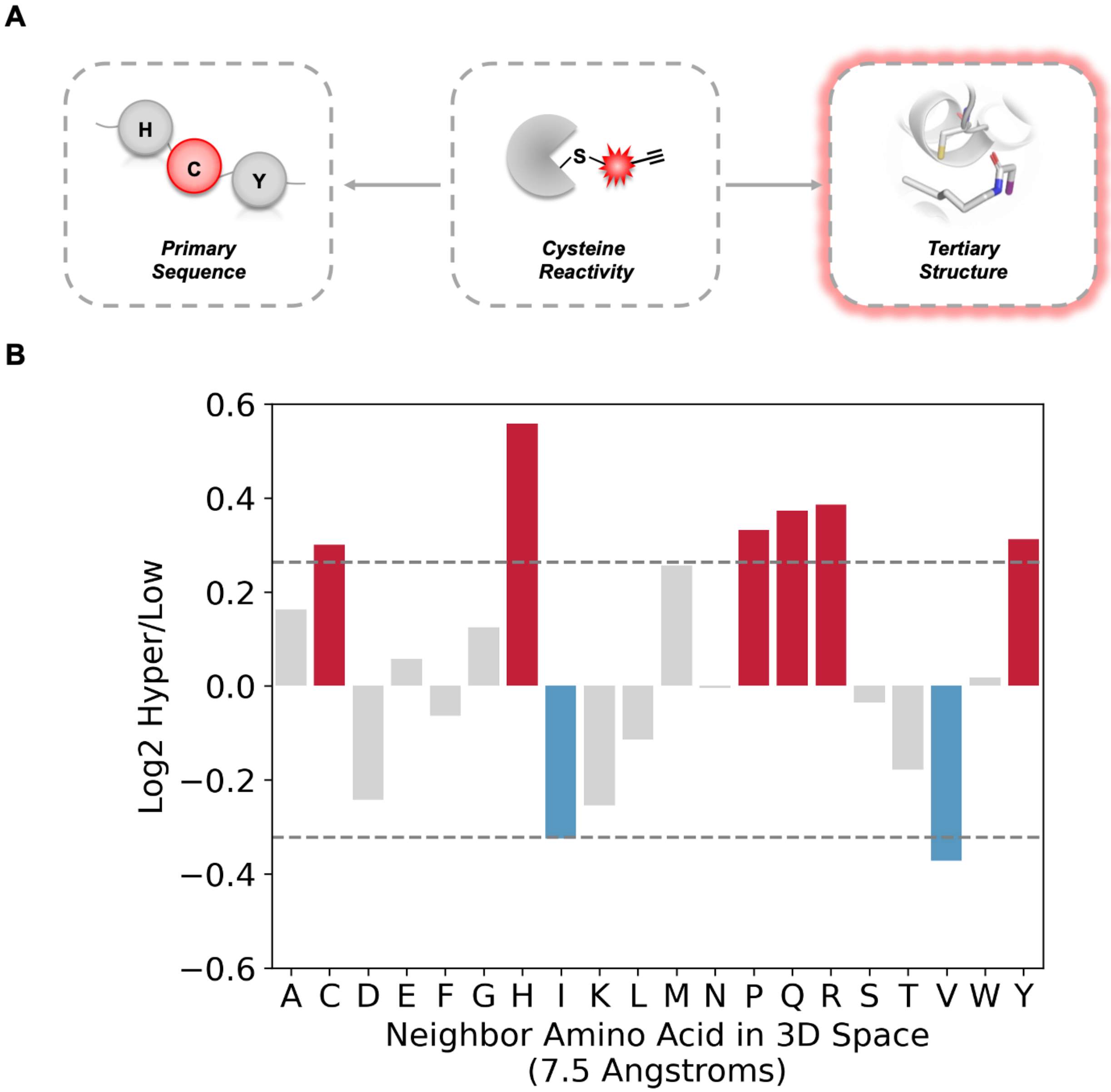


Figure 4. Amino acid content of IAA-reactive cysteines using 3D protein structures. (A) Overview of using tertiary structures of IAA-labeled cysteines resolved in associated PDB structures (306 high reactivity cysteines and 297 low reactivity cysteines). (B) Log_2 ratio of amino acid frequencies within a 7.5Å neighborhood around high reactivity cysteines relative to low reactivity cysteines. Red bars indicate enriched residues in high reactivity cysteine neighborhoods, while blue bars indicate depleted residues in these neighborhoods. See **Figure S7**, **Figure S8**, and **Data S3**.

Tertiary structure amino acid composition of high reactivity cysteines

With our curated set of structurally resolved cysteines in hand, we next sought to assess the amino acid content of IAA highly reactivity cysteine 3D neighborhoods (**Figure 4A**). Similar to our linear sequence analysis (**Figure 2**), we hypothesized that the 3D protein environment surrounding high reactivity cysteine residues should be enriched for reactivity-potentiating residues, such as histidine and cysteine. Therefore, to enable quantification of the proximal amino acid content around high and low reactivity cysteines, we aggregated the coordinates of all atoms within 7.5 Å of the sulfhydryl group (SG) atoms for each structurally resolved cysteine, excluding atoms from the cysteine residue itself. We selected 7.5 Å as a distance cutoff to ensure capture of neighboring residues without sampling more distal residues (**Figure S7**).

To prevent overcounting and to generate a non-redundant set of cysteine identifiers, residues were grouped by the corresponding PDB chain and residue number, retaining only unique residue identifiers (PDB_Chain_C#). The frequency of each amino acid within the high- and low reactivity cysteine neighborhoods was then calculated and normalized by the total number of unique residue identifiers within 7.5 Å of the SG atoms, accounting for potential differences, particularly for more buried cysteines.[17] To avoid overcounting, each residue was included only once if any of its atoms fell within the 7.5 Å radius, ensuring that residues were counted as unique entities rather than based on the total number of atoms they contributed.

This analysis identified a propensity of histidine and proline residues near high reactivity cysteines, aligning with our previous primary sequence analysis findings (**Figure 4B** and **Data S3**). Additionally, we observed an increase in arginine and glutamine residues and a decrease in hydrophobic residues, such as isoleucine and valine. Looking beyond these specific cysteine microenvironments, we observed generally similar amino acid content for proteins in our dataset compared to a UniProtKB reference human proteome (**Figure S8**), which indicates that our dataset is not inherently enriched or depleted for particular amino acids. Therefore, we conclude that the aforementioned high reactivity cysteine-specific amino acid enrichment represents bona fide features within the 3D cysteine microenvironment that drive cysteine nucleophilicity.

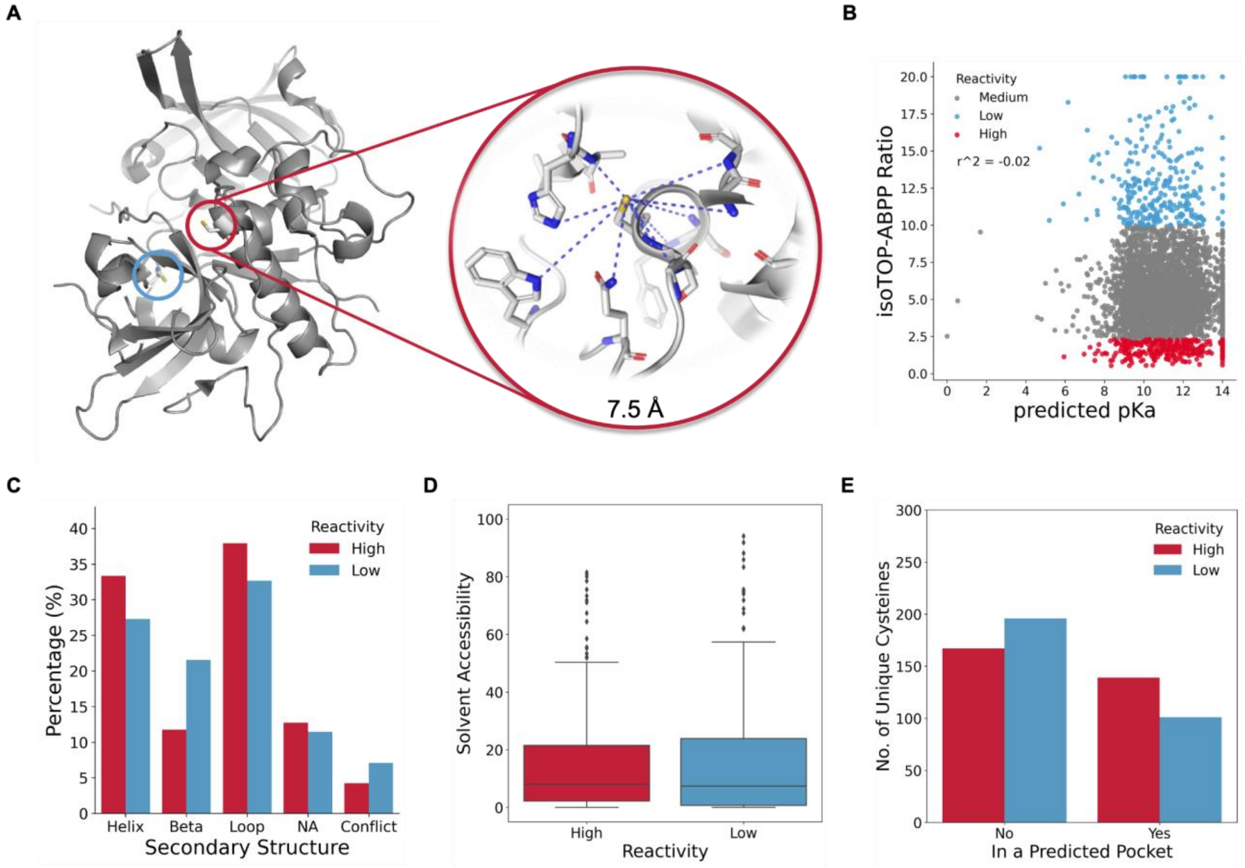


Figure 5. Chemical, reactivity and structural properties of IAA-reactive cysteines. (A) Collection of chemical, reactivity, and structural properties of IAA-reactive cysteines using tertiary structures. (B) Comparing computationally predicted pKa (PROPKA 3.1)[18] values and quantitative cysteine reactivity isoTOP-ABPP ratios ($R_{10:1}$). (C) Percentage of IAA-reactive cysteines in various secondary structure regions, as determined by DSSP-2.[38,39] (D) Comparison of computationally determined relative solvent accessibility (DSSP-2) and quantitative cysteine reactivity isoTOP-ABPP ratios ($R_{10:1}$). (E) Number of IAA-reactive cysteines in a predicted pocket (Fpocket 4.2)[40]. See **Figures S9-S14** and **Data S3**.

Descriptors of IAA-reactive cysteines from 3D structures

Building on our amino acid content analysis, we next extended this microenvironment analysis to generate a more comprehensive set of structural features for reactive cysteines. We aggregated descriptors in the following categories: residue proximity, general structural motifs, solvent accessibility, predicted pocket presence, predicted pKa metrics, overall amino acid content (AAC), amino acid interactions (AAI), hydrogen bond interactions, physicochemical properties.

We started with larger structural features, including secondary structure motifs and relative solvent accessibility (RSA) of cysteines, which we classified using the Dictionary of Secondary Structure-2[38,39] (DSSP-2). Parallel RSA values were also computed, based on the Kabsch and Sander method,[39] for cysteines resolved in PDB structures to assess their exposure within the associated crystal structure. Fpocket[40] release 4.2 was used to detect ligand-binding pockets and predicted pKa values were computed using PROPKA[18] v3.1. B-factor and disorder were assessed using BioPython[41] functions. We also analyzed amino acid physicochemical properties and hydrogen bond interactions, collecting 1D and 3D descriptors for residues within the 7.5 Å cutoff around cysteines. Amino acid type descriptors were then assigned based on residue and atom properties defined by Cheng et al,[42] with amino acid interaction descriptors assigned based on residue and atom properties defined by the Graph-based Residue neighborhood Strategy to Predict binding sites (GRaSP)[43] method. Hydrophobicity around the cysteine was evaluated using the Kyte-Doolittle[44] scale, and steric clashes were defined when the distance between the cysteine SG atom and a neighboring atom was less than the sum of their Van der Waals radii.[45] Hydrogen bond descriptors categorized neighboring atoms as donors or acceptors from backbone or side chains,[46–48] with counts divided by the total atoms within 5 Å and 7.5 Å distances, creating a weighted hydrogen bond profile. Rosetta[49] was used to compute energetic contributions of various physicochemical properties for each reactive cysteine, using talaris2013 weights. In total, we generated 82 features for each cysteine (**Data S3**). Full description of how the descriptors were generated can be found in the **Supplementary Computational Methods.**

pKa prediction is insufficient to predict cysteine reactivity

The availability of computational tools that predict cysteine pKa, most notably PROPKA[18], highlights a potential opportunity for rapid discovery of reactive cysteines. Therefore, in building our set of descriptors, we investigated whether PROPKA predictions of pKa could inform IAA reactivity—we acknowledge the clear limitation that IAA reactivity does not directly measure thiol pKa but instead provides a proxy for relative reactivity towards electrophiles. Towards understanding the relationship between pKa and IAA reactivity, we first examined five experimental cases where both reactivity and pKa had been directly measured (**Table S2**).[50–54] Several of these test cases corroborated the relationship between higher IAA reactivity and lower pKa values, such as C145 of MGMT which had a ratio of 0.87 and an experimental pKa of 5.3 (**Figure S9**).[50]

To further assess the relatedness of PROPKA predictions and measures of IAA-cysteine reactivity, we next expanded our analysis to all cysteines in our dataset. Our PROPKA analysis did not reveal a significant correlation between median theoretical predicted pKa values and isoTOP-ABPP reactivity measurements (**Figure 5B**). The average median predicted pKa for high reactivity cysteines was 11.18 versus 10.71 for low reactivity cysteines. Thus we conclude that PROPKA predicted pKa is generally not a useful proxy for IAA reactivity.

A small subset of both the high- and low reactivity cysteines had predicted pKa values that strongly contrasted with their measured reactivity. Exemplifying this difference, for the high reactivity cysteines, 34 residues had predicted pKa values greater than or equal to 14. For the low reactivity subset, 16 cysteines had predicted pKa values less than 8.5. Therefore, we opted to inspect these cysteines further so as to better understand the discrepancies between pKa prediction and measured IAA reactivity.

For the high reactivity subset, we noted 16 cysteines involved in disulfide bonds, as annotated by UniProtKB and resolved structurally. This category of cysteine is exemplified by the redox active disulfide between C32 and C35 of thioredoxin (TXN) (**Figure S10A**).[55] Five additional high reactivity cysteines were also likely localized to disulfide bonds, as indicated by the presence of a disulfide bond in at least one associated structure or when another cysteine sulfur atom was within 3 Å (**Figure S11** and **Data S3**). We also noted several additional proximal cysteine pairs just beyond this distance cutoff as exemplified by ATP-dependent RNA helicase, DDX3X, in which the sulfur atom of C317 is 5.1 Å away from the sulfur atom of C298 (**Figure S10B**).[56] Intriguingly and pointing to possible unique features of the low reactivity and low pKa prediction, only four of 16 cysteines were located in disulfide bonds, whereas six were located near zinc ions in their associated structures, including C166 of Hepatocyte growth factor-regulated tyrosine kinase substrate (HGS) and C150 of Zinc finger CCCH-type antiviral protein 1 (ZC3HAV1) (**Data S3**). Five of these cysteines also had UniProtKB annotations supporting their involvement in zinc binding or indicating their presence in zinc finger regions.Thus we conclude that the difference in reactivity and predicted pKa may stem from redox active disulfide bonds and metal coordination for the high reactivity and low reactivity cysteine subsets, respectively.

Prevalence of high reactivity cysteines in secondary structure motifs

Previous studies have suggested that high reactivity cysteines are often located near alpha-helices.[16] Therefore, we next investigated whether this enrichment held true for our newly generated descriptors. We used the DSSP-2 algorithm to classify cysteines into four main categories: helices, beta sheets, loops, and conflicting annotations (**Data S3**). Among these classifications, 102 high reactivity cysteines were found in helices, 36 in beta sheets, and 116 in loops. In comparison, we observed 81 low reactivity cysteines in helices, 64 in beta sheets, and 97 in loops (**Figure 5C**). As these analyses do not consider residue position in the secondary structure, we further subsetted the cysteines located in alpha helices to assess proximity to the helix N-terminus. We defined a cysteine as being near the N-terminus of a helix if the nitrogen atoms of the two downstream residues (i+2 and i+3) were part of a helix and within 5 Å, even if the cysteine itself was not located within the helix. With this added filtering, we observe an increased number of reactive cysteines at the N-terminus of helices relative to lowly reactive cysteines, which indicates that our data corroborates that of prior reports (**Figure S12**).

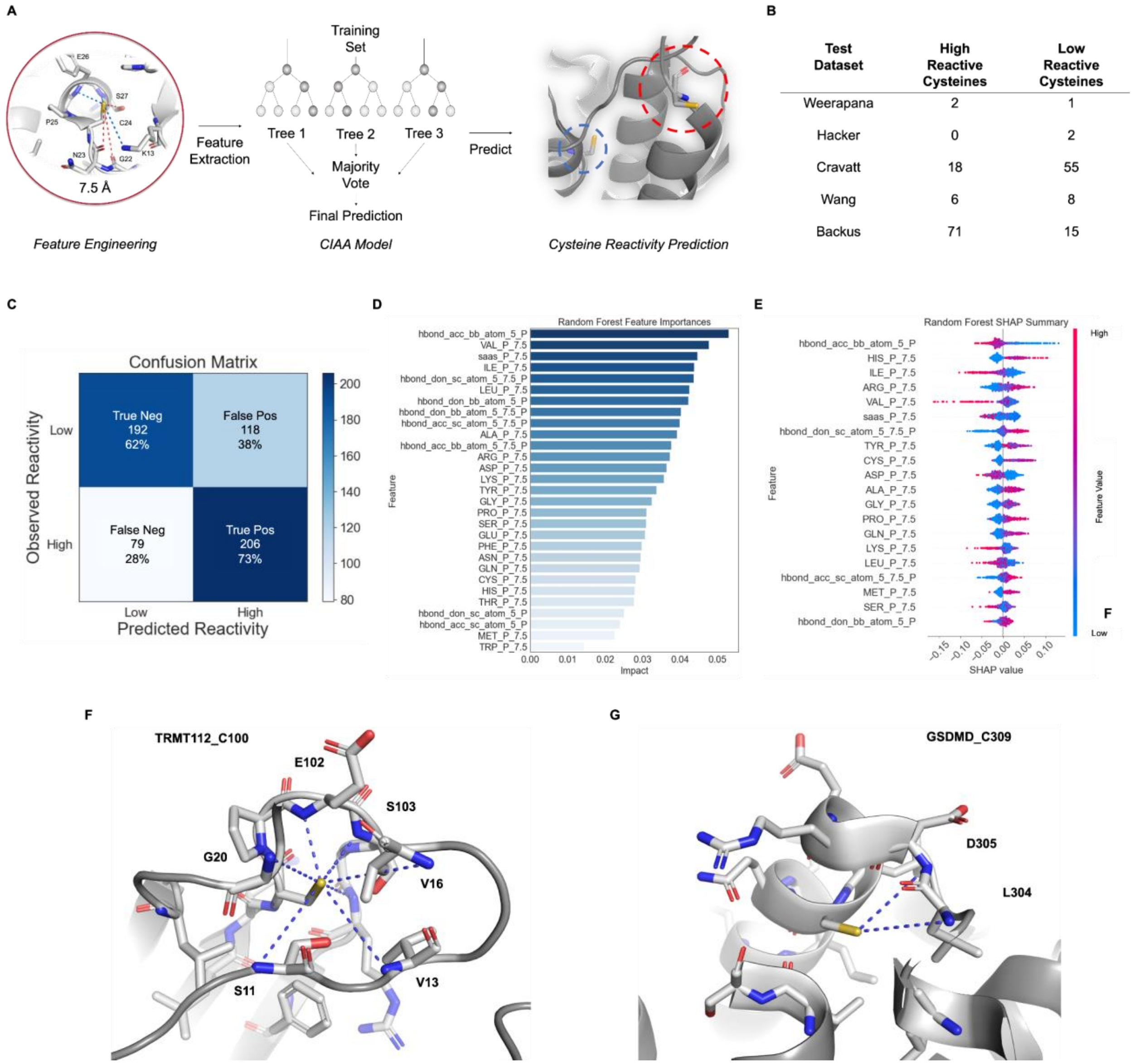
Relative solvent accessibility and pocket detection of high reactivity cysteines

We also examined the contribution of computationally predicted relative solvent accessibility (RSA) for each reactive cysteine. Again using DSSP-2 program, we calculated the median RSA for each structure associated with a UniProtKB_C# identifier. We did not identify a statistically significant difference in RSA between high- and low reactivity cysteines (Mann-Whitney U: 47,509.5, $p = 0.2970$) using either PDB structures or predicted protein structures from AlphaFold 2[57] (**Figure 5D** and **Figure S13**). On average, high reactivity cysteines had a median solvent accessibility of 15%, with 26% classified as “high-solvent accessible” (RSA ≥ 20), 15.7% as medium solvent accessible ($10 \leq \text{RSA} < 20$), and the remainder as low-solvent accessible. Thus we conclude that solvent accessibility is not sufficient to predict cysteine reactivity and that cysteines are frequently not highly solvent accessible, regardless of relative reactivity.

Given the relatively small nature of the cysteine SG, we postulated that solvent accessibility alone might inadequately capture the accessibility of specific residues to labeling with the comparatively bulky IAA probe. Therefore we also analyzed the proximity to pockets, using Fpocket[40] release 4.2. Consistent with our hypothesis, we found a modestly increased number of high reactivity cysteines were located in pockets, when compared to low reactivity residues, 45% (139 out of 306) versus 34% (101 out of 297), respectively (**Figure 5E**). These findings are consistent with prior studies which noted that cysteines identified by IA-DTB or liganded by scout fragments (e.g. KB02, KB03, or KB05) are typically not exclusively in highly exposed regions.[31,58,59] This pattern may reflect the functional importance of shielding high reactivity sites within potential binding pockets away from bulk solvent.

Correlation analysis of structural descriptors highlights the complex determinants of cysteine high reactivity

Guided by the suggestive enrichments for high reactivity cysteines in pockets and alpha helices, we next broadened our analysis to the rest of our descriptors, with the goal of pinpointing key features that drive cysteine reactivity. We assessed the correlation between the descriptors and experimental cysteine reactivity measurements via Pearson Correlation Coefficients (PCC). The highest PCC, 0.16, was observed for the percentage of hydrogen bond acceptor backbone atoms within a 5 Å radius of high reactivity cysteines (**Figure S14**). This inverse relationship suggests that less reactive cysteines may have fewer hydrogen bond donors available to stabilize the thiolate form. Unfortunately, no single descriptor emerged as a strong predictor of cysteine high reactivity. This lack of strong correlations between any individual descriptors and cysteine reactivity lead us to conclude that high cysteine reactivity towards IAA is likely governed by a combination of factors rather than any single structural feature.



Test Dataset	High Reactive Cysteines	Low Reactive Cysteines
Weerapana	2	1
Hacker	0	2
Cravatt	18	55
Wang	6	8
Backus	71	15

Figure 6. Features of reactive cysteines can be used to build CIAA, a random forest model to predict cysteine reactivity towards IAA. (A) Workflow of extracting features of cysteine reactivity using protein structures as input for a random forest algorithm to predict cysteines highly reactive and lowly reactive towards IAA. (B) Table of datasets obtained from literature, showing the number of randomly sampled unique highly-reactive and low reactivity cysteines used as input for our testing set. (C) Confusion matrix heatmap showing the distribution of true positive, false positive, true negative, and false negative cases from the random forest algorithm. The matrix provides a visual representation of the model's classification performance, where the rows represent the actual classes (high- or low reactivity) and the columns represent the predicted classes. The observed reactivity classes are based on quantitative cysteine reactivity isoTOP-ABPP ratios ($R_{10:1}$). (D) Bar graph showing the most important features of the model, where feature importance scores were calculated using Gini importance.[60] The height of each bar represents the relative contribution of each feature to the model's predictions, with higher bars

indicating greater importance in determining high- or low reactivity cysteines. (E) SHapley Additive exPlanations (SHAP) summary showing the impact of selected features on the predicted classification (high- or low reactivity cysteines).[61,62] Each point represents a test case, with the position on the x-axis indicating the magnitude and direction of the feature's effect on the prediction. The color of each point represents the feature value, with pink indicating higher feature values and blue indicating lower feature values. Features with larger SHAP values have a greater impact on the prediction. (F) Close up view of correctly predicted high reactivity C100 of Multifunctional methyltransferase subunit TRM112-like protein (TRNT112) (PDB: 6KHS). Hydrogens are omitted for clarity. Potential hydrogen bonds are represented by blue dashed lines. (G) Close up view of correctly predicted low reactivity C309 of Gasdermin-D (GSDMD) (PDB: 5NH1). Hydrogens are omitted for clarity. Potential hydrogen bonds are represented by blue dashed lines. See **Figure S15-23** and **Data S3**.

Supervised learning for initial model development

To test the hypothesis that a combination and features is driving cysteine reactivity, we set out to develop a model that could enhance our understanding of the structural drivers of cysteine reactivity (**Figure 6A**). Given the complexity of the data, we pursued a supervised machine learning approach to predict whether a cysteine was low reactivity (0) or high reactivity (1) towards IAA. Our goal was to identify patterns within the structural features that could distinguish between these two classes with a focus on correctly predicting the high reactivity cysteine class.

To maximize the number of high reactivity cysteines in our training set, we opted to use the entirety of our experimental dataset as the "ground truth." Therefore, to establish an external test dataset, we subjected several additional published cysteine reactivity datasets[11,26,63–65] to our curation pipeline, applying the same filtering criteria and structural processing as we had for our training set, ensuring consistency in data handling (**Data S2**). Our external test dataset contains a randomly sample set of unique cysteines not included in our training set (**Figure 6B**). Out of the proteins in the test set, 231 are shared with proteins in the training set, though their cysteine residues are distinct between the sets.

To determine the most suitable model for this task, we initially compared several machine learning algorithms, each offering distinct advantages based on the dataset's characteristics. We tested Random Forest (RF), K-Nearest Neighbors (KNN), Classification and Regression Tree (CART), Linear Discriminant Analysis (LDA), and Support Vector Machine (SVM) (**Figure S15A**). These algorithms were selected to cover a range of approaches, from ensemble methods (RF) to distance-based (KNN) and linear separation techniques (LDA and SVM), ensuring that we considered different ways of modeling the data. After running preliminary tests, we observed that while some algorithms excelled in certain aspects, they struggled with balancing the true positive rate (TPR) and false positive rate (FPR). To further optimize the models, we performed recursive feature elimination (RFE) (**Figure S15B**), which allowed us to reduce the feature set by selecting the most important descriptors. Despite these efforts, the best model performance we could achieve at this stage resulted in a TPR of 70% and an FPR of 44% (**Figure S15C**), testing on our external validation set.

Ablation studies to evaluate descriptor importance

We refined the model by conducting ablation studies that assessed the influence of different categories of descriptors on the model's TPR and FPR. By systematically removing individual descriptor categories, we identified the features that contributed most to true positive predictions (**Data S3**). This process revealed that including only three categories—AAC, hydrogen bond statistics, and RSA—resulted in a modest improvement for decreasing the false positive rate (baseline TPR of 71% and FPR of 39%) (**Figure 6C**). The optimized model comprised 29 features (**Figure S16**), with the most influential being the relative percentage of hydrogen bond acceptor backbone atoms within 5 Å, the percentage of valine residues within 7.5 Å, and RSA (**Figure 6D**).

During model optimization, we also observed a trend in cysteines represented by multiple PDB structures, which showed an increase in correct prediction rates compared to those with only one structure (**Figure S17**). For cysteines with a single structure, the correct prediction rate was 47%, while those with multiple structures achieved over 50% accuracy in 83 out of 267 cases. This indicates that additional structural data may provide further context for predictions.

Effect of multiple structural representations on prediction accuracy

For example, the CIAA method predicted C141 of Flap Endonuclease (FEN1) to be high reactivity in three out of four test structures, achieving 75% accuracy (**Data S3**). The structure 3Q8M, which yielded an incorrect prediction, included both Chains A and B, each bound to a double-stranded DNA segment, while the other structures—3Q8K, 5FV7, and 5ZOD—contained only one or no DNA segment (**Figure S18**). Studies show that FEN1's cap helices near C141 in α-helix 6 become more ordered upon DNA binding,[66] potentially altering access to C141 based on DNA-bound conformation.

Another example, C2093 in DNA-dependent protein kinase (DNA-PK), demonstrated a similar pattern where multiple structures captured conformation-induced dynamics upon DNA and Ku70/80 binding. The X-ray crystallography structure 5LUQ, representing Apo-DNA-PKcs, predicted C2093 as low reactivity. In contrast, the electron microscopy structures 6ZFP (DNA-PKcs "state 2"), 7OTY (DNA-PKcs), and 5Y3R (DNA-PK holoenzyme) each representing various conformational states, predicted C2093 as high reactivity. In these structures, DNA-PK undergoes conformational adjustments, such as rotations and flexing of the N-terminal arm toward the FAT domain,[67] altering the local environment around C2093 (**Figure S19**). These examples highlight how the inclusion of multiple structural states provides additional data that can influence predictive outcomes.

SHAP Analysis of feature contributions

To further explore the impact of these features, we performed a SHapley Additive exPlanations (SHAP) analysis. Shapley values, derived from cooperative game theory, quantify the average marginal contribution of each feature to the model's prediction of low reactivity (0) or high reactivity (1) cysteines.[62,68] In our analysis, positive SHAP values indicated an increased likelihood of predicting high reactivity cysteines, while negative values decreased this likelihood. Specifically, lower values of hydrogen bond acceptor backbone atoms within 5 Å, valine residues within 7.5 Å, and RSA were found to increase the model's ability to predict high reactivity cysteines (**Figure 6E**). These insights highlight the role of structural features related to hydrogen bonding, residue composition, and solvent accessibility as key drivers of the model's improved predictive performance.

Two examples of correctly predicted cases are high reactivity C100 in the Multifunctional methyltransferase subunit TRM112-like protein (TRNT112) and low reactivity C309 in Gasdermin-D (GSDMD) (**Figure 6F** and **Figure 6G**, respectively). The microenvironment of C100 features an abundance of backbone hydrogen bond donors from nearby residues such as Ser11, Gly20, and Ser103, compared to a limited number of backbone acceptor hydrogen bond atoms. In contrast, C309 in GSDMD has fewer hydrogen bond donors available in its microenvironment and is situated near the acidic residue Asp305.

Model limitations and performance across protein functional classes

It is important to acknowledge the limitations of our model by examining cases where it failed. We compared correct and incorrect predictions across experimental structure determination methods. The model performed consistently across methods, with the highest TPR of 69% for X-ray structures (n = 370 PDBs) and the lowest TPR of 50% for NMR structures (n = 31 PDBs) (**Figure S20**). Despite a true negative rate (TNR) of 72% for NMR structures, the false negative rate (FNR) was also 50%. Interestingly, many cysteines incorrectly predicted as high reactivity using NMR structures were from proteins involved in transcription or regulation, particularly DNA/RNA-binding proteins, which may undergo significant conformational changes upon ligand binding. Examples include C416 near the flexible loop region of Nucleus accumbens-associated protein 1 (NACC1)[69] and C1070 in the unstructured C-terminal region of bifunctional 3'-5' exonuclease/ATP-dependent helicase (WRN).[69,70] This suggests that including NMR structures from conformationally flexible proteins may have reduced model performance by introducing false negatives.

To explore whether protein functional classes influenced incorrect predictions, we further examined model accuracy across these classes. The model achieved the highest accuracy (78%) when predicting high reactivity cysteines in nucleic acid/small molecule-binding proteins, chaperones, transporters, channels, and receptors (**Figure S21**). However, it struggled with correctly predicting low reactivity cysteines for enzymes, leading to an increase in false positives. Most high reactivity cysteines in our training set fall within a modest RSA range (5-20%), but the model struggles with highly solvent-accessible cysteines in enzymes that appear ligandable but are not necessarily high reactivity (**Figure S22**). This could be due to missing descriptors that capture pre-binding states, hidden allosteric pockets, or metrics accounting for ligand accessibility and specific protein-ligand interactions. For instance, low reactivity C14 of Uroporphyrinogen-III synthase (UROS) ($R_{10:1}$ = 19.22) was incorrectly predicted to be high reactivity. However, C14 of UROS was shown to be liganded by an acrylamide derivative with a phenyl-oxazole substituent[71] and has a high RSA of 84% (PDB: 1JR2).

Model limitations and performance using AlphaFold 2 structures

We also explored the application of the CIAA model using AlphaFold 2 structures, as not all proteins in our experimental dataset had associated crystal structures in the PDB, or the reactive cysteines of interest were not resolved in their structures. AlphaFold 2[57] provides computational predictions of protein structures based on sequences for over 200 million proteins. Leveraging this abundance of data, we tested our CIAA model using AlphaFold 2 structures in place of PDB structures. We identified cases from our test set that were correctly predicted using PDB structures (**Figure 6C**) and obtained the corresponding AlphaFold 2 structures (accessed 2301). Upon testing, the model achieved an accuracy of 72.5% (**Figure S23A**).

Next, we examined whether we could use AlphaFold 2 structures to predict cysteine reactivity towards IAA for proteins lacking associated crystal structures in the PDB or those without resolved reactive cysteines. We identified such proteins from our experimental dataset (n = 409) and downloaded their AlphaFold 2 structures. However, unlike the prior performance with AlphaFold 2 structures, the model showed lower accuracy, achieving only 52.3%. Most of the misclassifications involved high reactivity cysteines incorrectly predicted as low reactivity (**Figure S23B**).

Discussion

To enhance the discovery of high reactivity and likely functional cysteine residues, here we developed "Cysteine reactivity towards IodoAcetamide Alkyne (CIAA)," an *in silico* method designed for high-throughput, high-coverage investigations of cysteine reactivity. CIAA incorporates published and in-house chemoproteomics studies, which in aggregate measure reactivity towards IAA for 9,783 cysteines, including 823 classified as high reactivity—thus our work more than doubles the number of known high reactivity cysteines previously reported in the literature. Enabled by this data, we mined protein structures to define features that indicate cysteine reactivity. Consistent with prior studies, we find that high reactivity cysteines are frequently located near histidines, prolines, and positively charged residues and are found in alpha helices.[13] Aligning with recent efforts to analyze a related class of ligandable, potentially "druggable," cysteines,[72] we also observe an enrichment for high reactivity cysteines in pockets—we expect that some of these residues could serve as useful starting points for drug development campaigns and that such highly reactive cysteines may prove particularly tractable for hit-to-lead optimization.

As none of these features alone were sufficient to provide a high confidence metric of cysteine likely reactivity, we incorporated all descriptors into a supervised random forest model, which resulted in an overall accuracy of 68%, with key predictive features including the depletion of hydrogen bond acceptor atoms, depletion of valine residues, and intermediate values of relative solvent accessibility. Although the model achieved a true positive rate of 71%, the false positive rate of 39% prompted further examination of its limitations. Many misclassified cysteines were located within conformationally dynamic proteins or highly solvent-accessible regions, indicating that protein dynamics, such as shifts between open and closed states, significantly impact reactivity predictions. For example, C285 of CASP1, experimentally classified as low reactivity, was predicted to be high reactivity by the CIAA model when analyzed in the active conformation of CASP1 (PDB: 6BZ9)—we expect this disconnect stems from the non-stimulated nature of the cellular proteomes analyzed, in which CASP1 should exist largely in the zymogen or inactive form. Thus we conclude that state-dependent cysteine reactivity may rationalize some of the differences observed between the model and proteomic measurements.[11,73] Looking beyond state-dependent activities, our work also highlights ongoing challenges in computational predictions, particularly in protein structure selection and dataset curation as being critical for model performance. Future efforts to improve our model's performance will likely benefit from incorporating protein dynamics and other state-specific features, such as protein interactions, together with stringent dataset and structure curation.

Looking beyond structurally resolved cysteines, the rapid growth of protein structure prediction, most notably via AlphaFold,[57] opens up tremendous opportunities for *in silico* discovery of reactive, functional, and ligandable cysteines proteome-wide in a species-agnostic manner. Our comparison of AlphaFold structures that either have or lack matched structures in the PDB reveals key differences in CIAA performance. For the former, the availability of matched structures resulted in high accuracy of the CIAA model. For the latter, performance was poor, likely due to larger dissimilarity between predicted structures and AlphaFold training data. We are optimistic that future implementations of AlphaFold 2 and related tools will prove compatible with *in silico* cysteine analysis, as this does not reflect an inherent limitation in predicted structures.[74,75] Such efforts will also benefit from ongoing efforts to increase chemoproteomic dataset size to further improve training set quality.[71,76–78]

Acknowledgements

This study was supported by a Beckman Young Investigator Award (K.M.B.), UCLA Jonsson Comprehensive Cancer Center Seed Grant (K.M.B.), DOD-Advanced Research Projects Agency (DARPA) D19AP00041 (K.M.B.), NIGMS System and Integrative Biology 5T32GM008185-33 (L.M.B.), and NIGMS R01 GM069832 (S.F.) We thank all members of the Houk, Forli, and Backus labs for helpful suggestions. This work used computational and storage services associated with the Hoffman2 Shared Cluster provided by UCLA Office of Advanced Research Computing's Research Technology Group.

Author Contributions

L.M.B., J.E., S.F. and K.M.B. conceived of the project. L.M.B. and F.S. collected data. L.M.B., J.E. and P.Y. performed data analysis. L.M.B and J.E. wrote software. S.F., J.E., M.H., and K.N.H. provided technical advice. L.M.B., S.F. and K.M.B. wrote the manuscript.

J.E. current address and affiliation: SIB Swiss Institute of Bioinformatics, University of Basel, Basel, Switzerland.

Conflicts of Interest

K.M.B. is a member of the advisory board at Matchpoint Therapeutics. All remaining authors declare no conflicts of interest.

Data Availability

The MS data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the dataset identifier PXD056064. File and peptide details are listed in **Data S1**.

Code Availability

Original code has been deposited at https://github.com/BackusLab/ciaa_app and is publicly available as of the date of publication.

Associated Content

Supplementary Methods and Supplementary Figures S1-S23 and Tables S1, S2 (PDF)

Data S1: Proteomic datasets and analyses related to Figure 1 and Figure 2 (xlsx)

Data S2: Datasets and analyses related to Figure 3 and Figure 4 (xlsx)

Data S3: Datasets and analyses related to Figure 5 and Figure 6 (xlsx)

References

1. Poole, L.B. (2015). The basics of thiols and cysteines in redox biology and chemistry. Free Radic. Biol. Med. *80*, 148–157. 10.1016/j.freeradbiomed.2014.11.013.
2. Go, Y.-M., Chandler, J.D., and Jones, D.P. (2015). The cysteine proteome. Free Radic. Biol. Med. *84*, 227–245. 10.1016/j.freeradbiomed.2015.03.022.
3. Walsh, C.T., Garneau-Tsodikova, S., and Gatto, G.J. (2005). Protein posttranslational modifications: the chemistry of proteome diversifications. Angew. Chem. Int. Ed *44*, 7342–7372. 10.1002/anie.200501023.
4. Moellering, R.E., and Cravatt, B.F. (2012). How chemoproteomics can enable drug discovery and development. Chem. Biol. *19*, 11–22. 10.1016/j.chembiol.2012.01.001.
5. Spradlin, J.N., Zhang, E., and Nomura, D.K. (2021). Reimagining druggability using chemoproteomic platforms. Acc. Chem. Res. *54*, 1801–1813. 10.1021/acs.accounts.1c00065.
6. 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. 10.1038/nature09472.
7. Nelson, J.W., and Creighton, T.E. (1994). Reactivity and ionization of the active site cysteine residues of DsbA, a protein required for disulfide bond formation in vivo. Biochemistry *33*, 5974–5983. 10.1021/bi00185a039.
8. Barglow, K.T., and Cravatt, B.F. (2007). Activity-based protein profiling for the functional annotation of enzymes. Nat. Methods *4*, 822–827. 10.1038/nmeth1092.
9. Kisty, E.A., Saart, E.C., and Weerapana, E. (2023). Identifying Redox-Sensitive Cysteine Residues in Mitochondria. Antioxidants (Basel) *12*. 10.3390/antiox12050992.
10. 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. 10.15252/msb.20209840.
11. Castellón, J.O., Ofori, S., Burton, N.R., Julio, A.R., Turmon, A.C., Armenta, E., Sandoval, C., Boatner, L.M., Takayoshi, E.E., Faragalla, M., et al. (2024). Chemoproteomics Identifies State-Dependent and Proteoform-Selective Caspase-2 Inhibitors. J. Am. Chem. Soc. *146*,

14972–14988. 10.1021/jacs.3c12240.

12. 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. 10.1002/cbic.202000870.

13. 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. 10.1021/acs.biochem.7b00897.

14. Sun, M.-A., Zhang, Q., Wang, Y., Ge, W., and Guo, D. (2016). Prediction of redox-sensitive cysteines using sequential distance and other sequence-based features. BMC Bioinformatics *17*, 316. 10.1186/s12859-016-1185-4.

15. Cao, J., and Xu, Y. (2024). Predicting cysteine reactivity changes upon phosphorylation using XGBoost. FEBS Open Bio *14*, 51–62. 10.1002/2211-5463.13737.

16. Fowler, N.J., Blanford, C.F., de Visser, S.P., and Warwicker, J. (2017). Features of reactive cysteines discovered through computation: from kinase inhibition to enrichment around protein degrons. Sci. Rep. *7*, 16338. 10.1038/s41598-017-15997-z.

17. Keßler, M., Wittig, I., Ackermann, J., and Koch, I. (2021). Prediction and analysis of redox-sensitive cysteines using machine learning and statistical methods. Biol. Chem. *402*, 925–935. 10.1515/hsz-2020-0321.

18. Olsson, M.H.M., Søndergaard, C.R., Rostkowski, M., and Jensen, J.H. (2011). PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pK Predictions. J. Chem. Theory Comput. *7*, 525–537. 10.1021/ct100578z.

19. Anandakrishnan, R., Aguilar, B., and Onufriev, A.V. (2012). H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. Nucleic Acids Res. *40*, W537-41. 10.1093/nar/gks375.

20. Soylu, İ., and Marino, S.M. (2016). Cy-preds: An algorithm and a web service for the analysis and prediction of cysteine reactivity. Proteins *84*, 278–291. 10.1002/prot.24978.

21. Harris, R.C., Liu, R., and Shen, J. (2020). Predicting Reactive Cysteines with Implicit-Solvent-Based Continuous Constant pH Molecular Dynamics in Amber. J. Chem. Theory Comput. *16*, 3689–3698. 10.1021/acs.jctc.0c00258.

22. Mapes, N.J., Rodriguez, C., Chowriappa, P., and Dua, S. (2019). Residue adjacency matrix based feature engineering for predicting cysteine reactivity in proteins. Comput. Struct. Biotechnol. J. *17*, 90–100. 10.1016/j.csbj.2018.12.005.

23. Nallapareddy, V., Bogam, S., Devarakonda, H., Paliwal, S., and Bandyopadhyay, D. (2021). DeepCys: Structure-based multiple cysteine function prediction method trained on deep neural network: Case study on domains of unknown functions belonging to COX2 domains. Proteins *89*, 745–761. 10.1002/prot.26056.

24. Gao, M., and Günther, S. (2023). HyperCys: A Structure- and Sequence-Based Predictor of Hyper-Reactive Druggable Cysteines. Int. J. Mol. Sci. *24*. 10.3390/ijms24065960.

25. 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. 10.1038/nature18002.

26. 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. 10.1016/j.cell.2020.07.001.

27. Shraga, A., Olshvang, E., Davidzohn, N., Khoshkenar, P., Germain, N., Shurrush, K., Carvalho, S., Avram, L., Albeck, S., Unger, T., et al. (2019). Covalent docking identifies a potent and selective MKK7 inhibitor. Cell Chem. Biol. *26*, 98-108.e5. 10.1016/j.chembiol.2018.10.011.

28. Lu, X., Smaill, J.B., Patterson, A.V., and Ding, K. (2022). Discovery of Cysteine-targeting Covalent Protein Kinase Inhibitors. J. Med. Chem. *65*, 58–83. 10.1021/acs.jmedchem.1c01719.

29. Amendola, G., Ettari, R., Previti, S., Di Chio, C., Messere, A., Di Maro, S., Hammerschmidt, S.J., Zimmer, C., Zimmermann, R.A., Schirmeister, T., et al. (2021). Lead Discovery of SARS-CoV-2 Main Protease Inhibitors through Covalent Docking-Based Virtual Screening. J. Chem. Inf. Model. *61*, 2062–2073. 10.1021/acs.jcim.1c00184.

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. 10.1016/j.chembiol.2023.04.004.

31. Takahashi, M., Chong, H.B., Zhang, S., Yang, T.-Y., Lazarov, M.J., Harry, S., Maynard, M., Hilbert, B., White, R.D., Murrey, H.E., et al. (2024). DrugMap: A quantitative pan-cancer analysis of cysteine ligandability. Cell *187*, 2536-2556.e30. 10.1016/j.cell.2024.03.027.

32. Hughes, C.S., Sorensen, P.H., and Morin, G.B. (2019). A Standardized and Reproducible Proteomics Protocol for Bottom-Up Quantitative Analysis of Protein Samples Using SP3 and Mass Spectrometry. Methods Mol. Biol. *1959*, 65–87. 10.1007/978-1-4939-9164-8_5.

33. Kong, A.T., Leprevost, F.V., Avtonomov, D.M., Mellacheruvu, D., and Nesvizhskii, A.I. (2017). MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. Nat. Methods *14*, 513–520. 10.1038/nmeth.4256.

34. Yu, F., Haynes, S.E., Teo, G.C., Avtonomov, D.M., Polasky, D.A., and Nesvizhskii, A.I. (2020). Fast Quantitative Analysis of timsTOF PASEF Data with MSFragger and IonQuant. Mol. Cell. Proteomics *19*, 1575–1585. 10.1074/mcp.TIR120.002048.

35. 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. 10.1038/nmeth.2646.

36. Chivers, P.T., Prehoda, K.E., and Raines, R.T. (1997). The CXXC motif: a rheostat in the active site. Biochemistry *36*, 4061–4066. 10.1021/bi9628580.

37. Wang, G., and Dunbrack, R.L. (2003). PISCES: a protein sequence culling server. Bioinformatics *19*, 1589–1591. 10.1093/bioinformatics/btg224.

38. Touw, W.G., Baakman, C., Black, J., te Beek, T.A.H., Krieger, E., Joosten, R.P., and Vriend, G. (2015). A series of PDB-related databanks for everyday needs. Nucleic Acids Res. *43*, D364-8. 10.1093/nar/gku1028.

39. Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers *22*, 2577–2637. 10.1002/bip.360221211.

40. Le Guilloux, V., Schmidtke, P., and Tuffery, P. (2009). Fpocket: an open source platform for ligand pocket detection. BMC Bioinformatics *10*, 168. 10.1186/1471-2105-10-168.

41. Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423. 10.1093/bioinformatics/btp163.

42. Cheng, S., Shi, T., Wang, X.-L., Liang, J., Wu, H., Xie, L., Li, Y., and Zhao, Y.-L. (2014). Features of S-nitrosylation based on statistical analysis and molecular dynamics simulation: cysteine acidity, surrounding basicity, steric hindrance and local flexibility. Mol. Biosyst. *10*, 2597–2606. 10.1039/c4mb00322e.

43. Santana, C.A., Silveira, S. de A., Moraes, J.P.A., Izidoro, S.C., de Melo-Minardi, R.C., Ribeiro, A.J.M., Tyzack, J.D., Borkakoti, N., and Thornton, J.M. (2020). GRaSP: a graph-based residue neighborhood strategy to predict binding sites. Bioinformatics *36*, i726–i734. 10.1093/bioinformatics/btaa805.

44. Kyte, J., and Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. *157*, 105–132. 10.1016/0022-2836(82)90515-0.

45. Bondi, A. (1964). van der Waals Volumes and Radii. J. Phys. Chem. *68*, 441–451. 10.1021/j100785a001.

46. Mazmanian, K., Sargsyan, K., Grauffel, C., Dudev, T., and Lim, C. (2016). Preferred Hydrogen-Bonding Partners of Cysteine: Implications for Regulating Cys Functions. J. Phys. Chem. B *120*, 10288–10296. 10.1021/acs.jpcb.6b08109.

47. Mundlapati, V.R., Ghosh, S., Bhattacherjee, A., Tiwari, P., and Biswal, H.S. (2015). Critical Assessment of the Strength of Hydrogen Bonds between the Sulfur Atom of Methionine/Cysteine and Backbone Amides in Proteins. J. Phys. Chem. Lett. *6*, 1385–1389. 10.1021/acs.jpclett.5b00491.

48. Roos, G., Foloppe, N., and Messens, J. (2013). Understanding the pK(a) of redox cysteines: the key role of hydrogen bonding. Antioxid. Redox Signal. *18*, 94–127. 10.1089/ars.2012.4521.

49. Alford, R.F., Leaver-Fay, A., Jeliazkov, J.R., O’Meara, M.J., DiMaio, F.P., Park, H., Shapovalov, M.V., Renfrew, P.D., Mulligan, V.K., Kappel, K., et al. (2017). The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J. Chem. Theory Comput.

13, 3031–3048. 10.1021/acs.jctc.7b00125.

50. Guengerich, F.P., Fang, Q., Liu, L., Hachey, D.L., and Pegg, A.E. (2003). O6-alkylguanine-DNA alkyltransferase: low pKa and high reactivity of cysteine 145. Biochemistry *42*, 10965–10970. 10.1021/bi034937z.

51. Witt, A.C., Lakshminarasimhan, M., Remington, B.C., Hasim, S., Pozharski, E., and Wilson, M.A. (2008). Cysteine pKa depression by a protonated glutamic acid in human DJ-1. Biochemistry *47*, 7430–7440. 10.1021/bi800282d.

52. Forman-Kay, J.D., Clore, G.M., and Gronenborn, A.M. (1992). Relationship between electrostatics and redox function in human thioredoxin: characterization of pH titration shifts using two-dimensional homo- and heteronuclear NMR. Biochemistry *31*, 3442–3452. 10.1021/bi00128a019.

53. Conway, M.E., and Harris, M. (2015). S-nitrosylation of the thioredoxin-like domains of protein disulfide isomerase and its role in neurodegenerative conditions. Front. Chem. *3*, 27. 10.3389/fchem.2015.00027.

54. Tolbert, B.S., Tajc, S.G., Webb, H., Snyder, J., Nielsen, J.E., Miller, B.L., and Basavappa, R. (2005). The active site cysteine of ubiquitin-conjugating enzymes has a significantly elevated pKa: functional implications. Biochemistry *44*, 16385–16391. 10.1021/bi0514459.

55. Cheng, Z., Zhang, J., Ballou, D.P., and Williams, C.H. (2011). Reactivity of thioredoxin as a protein thiol-disulfide oxidoreductase. Chem. Rev. *111*, 5768–5783. 10.1021/cr100006x.

56. Rosa E Silva, I., Smetana, J.H.C., and de Oliveira, J.F. (2024). A comprehensive review on DDX3X liquid phase condensation in health and neurodevelopmental disorders. Int. J. Biol. Macromol. *259*, 129330. 10.1016/j.ijbiomac.2024.129330.

57. 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. 10.1038/s41586-021-03819-2.

58. Offensperger, F., Tin, G., Duran-Frigola, M., Hahn, E., Dobner, S., Ende, C.W.A., Strohbach, J.W., Rukavina, A., Brennsteiner, V., Ogilvie, K., et al. (2024). Large-scale chemoproteomics expedites ligand discovery and predicts ligand behavior in cells. Science *384*, eadk5864. 10.1126/science.adk5864.

59. Biggs, G.S., Cawood, E.E., Vuorinen, A., McCarthy, W.J., Wilders, H., Riziotis, I.G., van der Zouwen, A.J., Pettinger, J., Nightingale, L., Chen, P., et al. (2024). Robust proteome profiling of cysteine-reactive fragments using label-free chemoproteomics. BioRxiv. 10.1101/2024.07.25.605137.

60. About us — scikit-learn 1.5.2 documentation https://scikit-learn.org/stable/about.html#citing-scikit-learn.

61. Rodríguez-Pérez, R., and Bajorath, J. (2020). Interpretation of Compound Activity Predictions from Complex Machine Learning Models Using Local Approximations and Shapley Values. J. Med. Chem. *63*, 8761–8777. 10.1021/acs.jmedchem.9b01101.

62. Wellawatte, G.P., Gandhi, H.A., Seshadri, A., and White, A.D. (2023). A perspective on explanations of molecular prediction models. J. Chem. Theory Comput. *19*, 2149–2160. 10.1021/acs.jctc.2c01235.

63. Bak, D.W., Pizzagalli, M.D., and Weerapana, E. (2017). Identifying functional cysteine residues in the mitochondria. ACS Chem. Biol. *12*, 947–957. 10.1021/acschembio.6b01074.

64. 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. *132*, 2851–2858. 10.1002/ange.201912075.

65. Yang, F., Chen, N., Wang, F., Jia, G., and Wang, C. (2022). Comparative reactivity profiling of cysteine-specific probes by chemoproteomics. Current Research in Chemical Biology *2*, 100024. 10.1016/j.crchbi.2022.100024.

66. Xu, H., Shi, R., Han, W., Cheng, J., Xu, X., Cheng, K., Wang, L., Tian, B., Zheng, L., Shen, B., et al. (2018). Structural basis of 5' flap recognition and protein-protein interactions of human flap endonuclease 1. Nucleic Acids Res. *46*, 11315–11325. 10.1093/nar/gky911.

67. Chen, X., Xu, X., Chen, Y., Cheung, J.C., Wang, H., Jiang, J., de Val, N., Fox, T., Gellert, M., and Yang, W. (2021). Structure of an activated DNA-PK and its implications for NHEJ. Mol. Cell *81*, 801-810.e3. 10.1016/j.molcel.2020.12.015.

68. Chen, H., Covert, I.C., Lundberg, S.M., and Lee, S.-I. (2023). Algorithms to estimate Shapley value feature attributions. Nat. Mach. Intell. 10.1038/s42256-023-00657-x.

69. Nakayama, N., Sakashita, G., Nagata, T., Kobayashi, N., Yoshida, H., Park, S.-Y., Nariai, Y., Kato, H., Obayashi, E., Nakayama, K., et al. (2020). Nucleus Accumbens-Associated Protein 1 Binds DNA Directly through the BEN Domain in a Sequence-Specific Manner. Biomedicines *8*. 10.3390/biomedicines8120608.

70. Perry, J.J.P., Yannone, S.M., Holden, L.G., Hitomi, C., Asaithamby, A., Han, S., Cooper, P.K., Chen, D.J., and Tainer, J.A. (2006). WRN exonuclease structure and molecular mechanism imply an editing role in DNA end processing. Nat. Struct. Mol. Biol. *13*, 414–422. 10.1038/nsmb1088.

71. 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. 10.1038/s41587-020-00778-3.

72. White, M.E.H., Gil, J., and Tate, E.W. (2023). Proteome-wide structural analysis identifies warhead- and coverage-specific biases in cysteine-focused chemoproteomics. Cell Chem. Biol. *30*, 828-838.e4. 10.1016/j.chembiol.2023.06.021.

73. Won, S.J., Zhang, Y., Reinhardt, C.J., Hargis, L.M., MacRae, N.S., DeMeester, K.E., Njomen, E., Remsberg, J.R., Melillo, B., Cravatt, B.F., et al. (2024). Redirecting the pioneering function of FOXA1 with covalent small molecules. Mol. Cell *84*, 4125-4141.e10. 10.1016/j.molcel.2024.09.024.

74. Holcomb, M., Chang, Y.-T., Goodsell, D.S., and Forli, S. (2023). Evaluation of AlphaFold2

structures as docking targets. Protein Sci. *32*, e4530. 10.1002/pro.4530.

75. Fournier, Q., Vernon, R.M., van der Sloot, A., Schulz, B., Chandar, S., and Langmead, C.J. (2024). Protein language models: is scaling necessary? BioRxiv. 10.1101/2024.09.23.614603.

76. Shikwana, F., Heydari, B., Ofori, S., Truong, C., Turmon, A., Darrouj, J., Holoidovsky, L., Gustafson, J., and Backus, K. (2024). CySP3-96 enables scalable, streamlined, and low-cost sample preparation for cysteine chemoproteomic applications. 10.26434/chemrxiv-2024-jm4n0.

77. Burton, N.R., and Backus, K.M. (2024). Functionalizing tandem mass tags for streamlining click-based quantitative chemoproteomics. Commun. Chem. *7*, 80. 10.1038/s42004-024-01162-x.

78. Yang, K., Whitehouse, R.L., Dawson, S.L., Zhang, L., Martin, J.G., Johnson, D.S., Paulo, J.A., Gygi, S.P., and Yu, Q. (2024). Accelerating multiplexed profiling of protein-ligand interactions: High-throughput plate-based reactive cysteine profiling with minimal input. Cell Chem. Biol. *31*, 565-576.e4. 10.1016/j.chembiol.2023.11.015.