Title

Deciphering the Therapeutic Accessibility of the Human Cysteinome using Experimental Quantitative Chemoproteomics

Author

Boatner, Lisa Marie

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Deciphering the Therapeutic Accessibility of the Human Cysteinome using Experimental Quantitative Chemoproteomics

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy in Chemistry

by

Lisa Marie Boatner

2024

ABSTRACT OF THE DISSERTATION

Deciphering the Therapeutic Accessibility of the Human Cysteinome using Experimental Quantitative Chemoproteomics

by

Lisa Marie Boatner
Doctor of Philosophy in Chemistry
University of California, Los Angeles, 2024
Professor Keriann M. Backus, Chair

Small molecule chemical probes are valuable tools for modulating protein function and have the potential to serve as leads for future medications. However, the pharmacological targeting of the human proteome with FDA-approved small molecules remains limited, addressing only 4% of all proteins. Furthermore, ~80% of proteins lack well-defined binding pockets for engagement by conventional small drug-like molecules. Mass spectrometry-based cysteine chemoproteomics has emerged as a promising strategy to bridge this druggability gap by mapping cysteine 'druggability' across the proteome. However, key challenges persist, including limited sampling (~13% of all cysteines), insufficient stratification of functional significance, and limited mechanistic insights into the labeling preferences of electrophilic compounds.

This work integrates experimental and computational approaches to address these challenges and improve the design and analysis of cysteine chemoproteomics datasets. First, the Mass Spectrometry-based Chemoproteomics Detected Amino Acids (MS-CpDAA) Analysis Suite was developed to streamline the deconvolution of covalent labeling sites from high-throughput chemoproteomics experiments and to quantify the performance of novel experimental methods for expanding cysteine coverage (Chapter 1). Using MS-CpDAA, we expanded cysteine coverage 5.5-fold compared to prior studies, identifying 34,225 covalently labeled cysteines. Building on this, CysDB, a publicly accessible SQL database with an interactive web interface, was established to aggregate experimental measures of cysteine reactivity alongside structural and functional annotations for over 24% of the cysteinome (Chapter 2). Designed to integrate diverse datasets and prioritize protein targets, CysDB provides a scalable platform for advancing the field. Designed to facilitate target prioritization, CysDB also provides a scalable platform for data integration and supports continued learning as the field evolves. Finally, CIAA (Cysteine reactivity towards IodoAcetamide Alkyne), a random forest model, was developed to predict cysteines with enhanced reactivity toward the small molecule iodoacetamide alkyne (IAA) (Chapter 3). CIAA offers a structure-based approach to investigating protein-ligand interactions, linking cysteine reactivity to druggability and functionality.

Together, this dissertation expands our understanding of the druggable cysteinome by providing computational resources and methodologies to target biologically significant proteins previously considered 'undruggable' and advancing approaches for covalent drug design. Furthermore, these approaches can be readily adapted to assess the druggability of other residues, such as lysines and tyrosines, across the human proteome. By addressing key challenges in cysteine chemoproteomics, these approaches contribute to a broader foundation for structure-based investigations of protein functionality and ligandability, offering valuable contributions to the fields of drug discovery and precision medicine.

This dissertation of Lisa Marie Boatner is approved.

Kenneth L. Lange

Joseph A. Loo

Kendall N. Houk

Keriann M. Backus, Committee Chair

University of California, Los Angeles

2024

DEDICATION

This dissertation is dedicated to my parents,

David M. Boatner and Mary C. Goudreault.

I love you to the moon and back.

"Don't underestimate me! I don't quit and I don't run."

– Uzumaki Naruto

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 2:

Chapter 3:

LIST OF TABLES

Chapter 3:

ACKNOWLEDGEMENTS

“Stand up and walk. Keep moving forward.

You’ve got two good legs. So get up and use them.

You’re strong enough to make your own path.”

– Edward Elric

Science, like life, is a team effort. I am deeply grateful to everyone who has supported me throughout this journey, as I could not have achieved such growth without your encouragement and belief in me.

First, I would like to thank my Ph.D. advisors, Dr. Keriann Backus and Dr. Kendall Houk. Dr. Backus is one of the most passionate scientists I have encountered, especially in the field of chemoproteomics, and her enthusiasm for science has been both inspiring and motivating. Dr. Houk is one of the sharpest minds I have ever met, thank you for bringing a sense of joy and optimism to every interaction. I have valued the feedback from both of you, which has helped me improve my technical expertise, communication skills, and confidence in expressing my ideas. I would also like to extend my gratitude to my committee members, Dr. Kenneth Lange and Dr. Joseph Loo, for their openness and willingness to meet with me to discuss scientific concepts and career advice. Their guidance has been invaluable throughout this process. Dr. Loo, your lightheartedness and humor have made the challenges much easier to navigate.

Second, I would like to thank my collaborators, Dr. Stefano Forli, Jerome Eberhardt, Dr. Devin Schweppe, and Dr. Alexey Nesvizhskii, for their invaluable support and contributions. Dr. Forli, your critical feedback and openness to scientific arguments have been both challenging and rewarding. Jerome, thank you for going out of your way to help me during the early stages of my work. Dr. Schweppe and Dr. Nesvizhskii, your kindness and encouragement have made conferences and collaborations all the more enjoyable.

Third, I would like to thank my past and present mentors who gave me the time and opportunities to grow. Rachel Garner, thank you for seeing something in me even when I couldn't see it in myself. Beyond being my dance instructor, you taught me valuable life lessons that I carry with me to this day and that have profoundly influenced my perseverance. You instilled in me a strong sense of discipline and commitment—I still remember staying in the studio after our one-on-one lessons to practice everything you told me to work on. You also provided a safe space and the encouragement to embrace self-expression, for which I will always be grateful.

Dr. Pedro Metola, thank you for approaching me one day in Welch Hall while I was jamming out and working on my general chemistry homework to ask if I would be interested in undergraduate research. Little did we know that simple interaction would lead to my joining the Crooks lab—a pivotal moment in my academic journey.

Dr. Richard Crooks, thank you for supporting my future goals, even after I graduated from undergrad. It was in your lab, under your guidance, that I learned how challenging yet essential the process of "troubleshooting" truly is. Your mentorship during that time has profoundly shaped my approach to problem-solving and research in countless ways.

Fourth, I would like to thank all past and present members of the Backus and Houk groups for making the challenges of lab life not only bearable but truly rewarding. The members of the Cysteine Chapel are incredibly hardworking, exceptionally caring, and full of chaotic energy and camaraderie that have made every day in the lab fun! It has been a true pleasure to connect with them—whether through spontaneous coffee or boba runs, lunchtime cookies and coffee, movie nights, holiday celebrations, or even attending the many beautiful lab weddings.

Similarly, I have immense appreciation for the Houkie "tiptorial" presentations and workshops, as well as the Houkie Party Planning Committee. Despite members being spread across the world, the Houk group always finds ways to come together, whether to celebrate birthdays, farewells, or life's big moments. These gatherings exemplify the spirit of scientific collaboration and friendship. May we continue to "push back the frontiers of science!"

Fifth, I would like to thank my friends who have been too generous to me over the last few years. Molly, you were one of my biggest cheerleaders at UT Austin. Even though we faced some difficult moments, I have always admired your unwavering determination and the fierceness with which you tackled every challenge. You have been a true role model to me, teaching me the importance of standing up for myself as I watched you do the same with strength and grace.

Merin, I can't believe it has been almost a decade since we met. I am endlessly grateful for our friendship and for how deeply you've been there for me throughout the years—whether it's answering my calls to let me vent, sending emotive voice messages, or spending hours on phone or video calls while we do chores. Thank you for making time to visit me during grad school, even while you were juggling medical school. Somehow, you always know exactly what to say, your excitement and happiness for me, even in moments when I didn't realize I needed it, have made such a difference in my life. Words can't fully express how much your friendship means to me.

Ernest, Ryan, and Phillip—thank you for staying connected over the years. Ernest and Ryan, I will always cherish our Mammoth trip, visiting Mochi on the East side, and bonding over our love of food and travel. Phillip, I love our phone chats about science, life, and anime, and I genuinely appreciate you joining me to volunteer for suicide prevention.

Maria, José, Sunny, and Heta—thank you for always being my rocks during graduate school. I will forever treasure our fancy lunches, hangouts at the Culver City Mall, and Mario Kart nights, and Universal/Disney trips!

Dr. Maria Palafox, you are my Longhorn soul sister. I always appreciated you popping over to check on me, and I love how supportive our friendship is. Your words of affirmation on low days have lifted me up countless times, and I'm so grateful for your constant encouragement.

Soon-to-be Dr. José Castellón, you have an incredible ability to make everyone feel included and take the time to check in with those around you. It's something I really admire about you. Also, thank you for holding my hand (literally and figuratively) while I learned how to do a Western blot.

Dr. Tianyang Yan (aka Sunny), your nickname couldn't suit you better! Beyond your sunny personality, I've loved how our friendship has grown. I can always count on you to text back when I need to vent or share good news. Your animated and competitive spirit makes every outing, whether it's playing DDR at Dave & Buster's or simply taking a break from grad school—a true joy.

Sixth, I would like to thank my housemates at 2123 ½ Burnside—the Burnside crew. I couldn't have done this without you: original residents Heta, Nik, and Billy, and new residents Kelsey and Cody. Thank you for the fun banter, four-hour Monopoly sessions on Thanksgiving, and hosting the best Halloween parties. Together, we've shared amazing and gloriously bad ideas, and I wouldn't trade those memories for anything.

Dr. Heta Desai, who would have thought teaching me isoTOP would lead to so many adventures? You're the ultimate partner in crime, always up for any side quest—including agreeing to live with me! I still can't believe you let me cut 10 inches off your hair for charity during the pandemic. Living with you has been such a joy, from crafting and jazz concerts to driveway dance sessions. I can't wait to create more memories together in Boston!

Soon-to-be Dr. Nikolas Burton, thank you for resetting the microwave timer, traveling to national parks, and indulging my existential musings about life—even when you were just trying to enjoy your ice cream from Saffron and Rose. You're the best driver and one of the most reliable people I know—even if we ran out of gas while you were driving my car! I couldn't have summited El Capitan without you and loved our late-night REI run before our 16-mile hike.

Dr. Joseph Treacy, thank you for entertaining my endless "why" questions, rescuing me when I locked my keys in my car after hiking Mt. Baldy, and persistently encouraging me to keep going and step outside my comfort zone. You're one of the most hard-working and driven people I know, and I'm grateful for your support—whether tackling challenges together or sharing unforgettable (and tear-inducing) experiences, like attempting the spiciest level of Howlin' Ray's hot chicken!

Seventh, I want to thank my family. While there are so many Boatners (and Boatner-adjacent family members) I need to thank, I want to especially acknowledge David, Mary, Cooper, and Mila. Dad, thank you for patiently listening to me practice my oral presentations multiple times a day—and on several occasions. Mom, thank you for always answering my late-night phone calls and sending me the perfect words to hype me up. Cooper, thank you for being there to comfort me on my toughest days. And Mila, thank you for being the light in my life. Your unwavering love and support have meant the world to me. I love you all so much and am endlessly grateful for having you as my biggest advocates.

Eighth, I want to thank the city of Los Angeles and the incredible communities I've been fortunate to be part of. From the vibrant hiking groups that connected me to the beauty of SoCal mountains, to the inspiring organizations like AFSP, NAMI, and CASA, I've found a sense of purpose and connection that has enriched my life beyond the lab. These experiences have taught me resilience, compassion, and the value of giving back, all of which have shaped my character in profound ways.

Last but not least, I also want to thank the SoCal mountains—specifically Cucamonga Peak, Mt. San Jacinto, Mt. Baldy, and Mt. Whitney. No one has witnessed the ups and downs of my journey more than those trails. Hiking these mountains has given me the space to think through scientific ideas, plan manuscripts, grants, and presentations, and reflect during moments of both joy and hardship. They've been there when I've celebrated, strategized, and even cried, offering clarity and resilience in ways I'll forever cherish.

Edward Elric's words perfectly encapsulate the perseverance, support, and determination that have carried me throughout graduate school. From the wisdom of mentors and the encouragement of loved ones to the clarity found on mountain trails, every step has been guided by the belief that I am strong enough to forge my own path. I owe so much of that strength to the incredible people and experiences I've been fortunate to have along the way. Thank you all for helping me find my footing and for walking this path with me.

Chapter 2 of this dissertation is a version of a published manuscript: Boatner, L. M.; Palafox, M. F.; Schweppe, D. K.; Backus, K. M. CysDB: a human cysteine database based on experimental quantitative chemoproteomics. *Cell Chemical Biology* **2023**, *30*(6), 683-698. https://doi.org/10.1016/j.chembiol.2023.04.004. Reprinted with permission. Copyright © 2023 Elsevier Ltd. L.M.B., D.K.S., and K.M.B. conceived the project. L.M.B. and M.F.P. performed data analysis. L.M.B. wrote software and created the database. D.K.S. provided technical advice. L.M.B. and K.M.B. wrote the manuscript.

Chapter 3 of this dissertation is a version of a manuscript in preparation: Boatner, L. M.; Eberhardt, J.; Shikwana, F.; Lee, P.; Houk, K. N.; Forli, S.; Backus, K. M. CIAA: Integrated Proteomics and Structural Modeling for Understanding Cysteine Reactivity with Iodoacetamide Alkyne. *Manuscript in preparation* 2024. L.M.B., S.F. and K.M.B. conceived the project. L.M.B. and F.S. collected data. L.M.B., J.E., and P.L. performed data analysis. L.M.B. and J.E. wrote software. S.F., J.E. and K.N.H. provided technical advice. L.M.B. and K.M.B. wrote the manuscript.

VITA

EDUCATION

University of California, Los Angles August 2019 – December 2024

Master of Science in Chemistry

The University of Texas at Austin August 2014 – December 2019

Bachelor of Science and Arts in Chemistry

Bachelor of Science in Computational Biology

Elements of Computing Certificate

Business Foundations Certificate

SELECTED PUBLICATIONS

1. **Boatner, L. M.***; Eberhardt, J.; Shikwana, F.; Lee, P.; Houk, K. N.; Forli, S.; Backus, K. M. CIAA: Integrated Proteomics and Structural Modeling for Understanding Cysteine Reactivity with Iodoacetamide Alkyne. *In Preparation* **2024**.
2. Desai, H.; Andrews, K. H.; Bergersen, K. V.; Ofori, S.; Yu, F.; Shikwana, F.; Arbing, MA.; **Boatner, L. M.***; Villanueva, M.; Ung, N.; Reed, E. F.; Nesvizhskii, A. I.; Backus, K. M. Chemoproteogenomic stratification of the missense variant cysteinome. *Nature Communications* **2024**, *15*(1), 9284.
3. **Boatner, L. M.***; Palafox, M. F.; Schweppe, D. K.; Backus, K. M. CysDB: A Human Cysteine Database based on Experimental Quantitative Chemoproteomics. *Cell chemical biology* **2023**, *30*, 683–698. e3.https://doi.org/10.1016/j.chembiol.2023.04.004.
4. Tang, K. C.; Cao, J.; **Boatner, L. M.***; Li, L.; Farhi, J.; Houk, K. N.; Spangle, J.; Backus, K. M.; Raj, M. Tunable Amine-Reactive Electrophiles for Selective Profiling of Lysine. *Angewandte Chemie* **2022**, *134*(5), e202112107.

5. Yan, T.; Desai, H. S.; **Boatner, L. M.***; Yen, S. L., Cao, J.; Palafox, M. F.; Jami-Alahmadi, Y.; Backus, K. M., SP3-FAIMS chemoproteomics for high coverage profiling of the human cysteinome. *Chembiochem* **2021**, *22*(10), 1841-1851.
6. Cao, J.; Armenta, E.; **Boatner, L. M.***; Desai, H.S.; Burton, N.R.; Armenta, E.; Chan, N.J.; Castellón, J.O.; Backus, K. M. Multiplexed CuAAC Suzuki–Miyaura Labeling for Tandem Activity-Based Chemoproteomic Profiling. *Analytical Chemistry* **2021**, *93*(4), 2610-2618.
7. Walgama, C.; Nguyen, M. P.; **Boatner, L. M.***; Richards, I.; Crooks, R. M. Hybrid paper and 3D-printed microfluidic device for electrochemical detection of Ag nanoparticle labels. *Lab on a Chip* **2020**, *20*(9), 1648-1657.

SELECTED PRESENTATIONS

1. 2024 19th Annual Drug Discovery Chemistry (DDC) Conference, San Diego, CA. "Unraveling Cysteine Reactivity for Therapeutic Targeting by Integrating Chemoproteomics and Computational Sciences."
2. 2023 18th Annual US Human Proteome Organization (HUPO) Conference, Chicago, IL. "Deciphering the Therapeutic Accessibility of the Human Cysteinome."

SELECTED AWARDS

1. 2024 Charles J. Pederson Dissertation Award, UCLA Chemistry Department
2. 2024 Rising Star in Computational Mass Spectrometry, Females in Mass Spectrometry (FeMS), OpenMS, University of Toronto and Chan Zuckerberg Initiative (CZI)
3. 2022 Excellence in Research Award, UCLA Chemistry Department
4. 2021 Graduate Research Fellowship Program Honorable Mention, National Science Foundation (NSF)
5. 2020 Systems and Integrative Biology Training Grant, National Institute of Health (NIH)

Chapter 1: Introduction

One of the biggest challenges in chemical biology, particularly in drug development, is addressing the "druggability gap." Out of the 20,000 protein-coding genes in the human genome, over 3,000 are implicated in disease.[1–4] Yet, fewer than 4% of human proteins have been successfully targeted by FDA-approved small molecule drugs,[5] demonstrating the difficulty of translating genetic insights into pharmaceutical advancements.

This challenge is compounded by the fact that most human proteins lack chemical probes capable of selectively modulating their activity, leaving entire classes historically labeled as "undruggable" or, more recently, "tough-to-drug." While classically druggable proteins, such as enzymes (CASP3, TXN, and DUBs)[7–9] and receptors (BTK and EGFR),[10–13] possess well-defined, hydrophobic, solvent-accessible binding pockets suitable for engagement by small, drug-like molecules, undruggable proteins often lack these features. These include transcription factors, intrinsically disordered proteins, and proteins involved in complex protein-protein interactions (PPIs), such as STAT3, NF-kB, p53, and c-Myc.[6,14–17] Despite their structural and functional challenges, thousands of genetic drivers of human disease have been discovered within the undruggable proteome, representing a pool of novel potential therapeutic targets.[18–20]

To overcome the limitations of traditional therapeutic strategies for undruggable proteins, innovative approaches are needed. One promising strategy involves leveraging the unique chemical properties of cysteine residues, which play critical roles in protein function and present new opportunities for therapeutic targeting. Although cysteine is one of the rarest amino acids (comprising only 2.3% of the human proteome),[21] it is highly conserved due to its functional significance. Cysteines serve as redox sensors, catalytic nucleophiles, structural motifs, and targets for chemical probes and FDA-approved drugs.[22–26] Cysteine-reactive compounds, particularly covalent inhibitors, have emerged as a promising class of molecules for modulating protein function, especially for tough-to-drug proteins.[27,28] A notable example is the successful targeting of KRAS, a protein previously thought to be undruggable, through cysteine-reactive compounds that label the Gly12Cys mutant form of KRAS.[29,30] Consequently, the identification of functional and potentially druggable cysteines is a central challenge of functional biology and drug development.

Identifying functional and potentially druggable cysteines requires robust tools to profile their reactivity and "druggability" (termed "ligandability") across the proteome. Mass spectrometry-based chemoproteomics has emerged as a powerful approach for this purpose, enabling systematic mapping of cysteine reactivity and ligandability. One such method, Isotopic Tandem Orthogonal Proteolysis-Activity-Based Protein Profiling (isoTOP-ABPP), has been applied to target cysteine residues using a pan-cysteine reactive probe, iodoacetamide alkyne (IAA).[31] IAA contains an alkyne group, which facilitates copper-catalyzed azide-alkyne cycloaddition (CuAAC), commonly known as "click" chemistry.

A general chemoproteomics workflow begins by treating lysates or cells with micromolar concentrations of IAA or a cysteine-reactive electrophile functionalized compound. After labeling, peptides are conjugated with an enrichment tag via click chemistry, selectively enriched using affinity purification (e.g., streptavidin), and identified by liquid chromatography-tandem mass spectrometry (LC-MS/MS). Biotinylated peptides are first ionized and detected at the MS1 level, where their mass-to-charge (m/z) ratios are recorded. These precursor ions are then isolated and fragmented at the MS2 level, allowing for detailed sequencing and identification. Peptide sequences are typically determined using data-dependent acquisition (DDA), where experimental spectra are matched to reference spectra from a specified database or FASTA file, enabling identification of both the peptide sequence and the modification site targeted by the IAA probe.[32]

The application of isoTOP-ABPP has enabled proteome-wide measurements that have identified numerous potentially druggable sites. Measurements of cysteine reactivity have revealed hundreds of highly reactive cysteine residues within both known functional sites and previously unannotated protein domains.[31,33,34] In 2016, Backus et al. identified over 600 proteins liganded by cysteine-reactive compounds, with only 27 of these proteins overlapping with targets of FDA-approved drugs, highlighting the potential of cysteine chemoproteomics to access new "druggable space." A key limitation of current chemoproteomic profiles, however, is their ability to assay only a small fraction (~13%) of the approximately 260,000 cysteines present in the human proteome.[35,38]

Over the past decade, significant advancements in cysteine chemoproteomics have greatly expanded the field's capabilities. Key developments include enhanced sample enrichment methods, such as single-pot, solid-phase-enhanced sample preparation (SP3), simplified and ultrafast peptide enrichment and release (superTOP-ABPP), and desthiobiotin-based tags (isoDTB), which improve the capture and identification of labeled cysteines.[36–40] Quantitative labeling strategies, including isobaric tags, tandem mass tags (TMT) and 96-well plate assays, now allow for higher throughput and simultaneous profiling across multiple conditions.[41–51] Additionally, innovations in cysteine-reactive electrophiles—such as maleimides, heteroaromatic sulfones, bifunctional probes, and stereoprobes—have increased the precision and versatility of cysteine-targeting warheads.[52–62]

Mass spectrometry has seen remarkable advancements in data acquisition strategies and instrumentation, significantly enhancing the analytical sensitivity and depth of proteomics studies. Technologies such as data-independent acquisition (DIA) methods,[63–66] the Orbitrap Ascend

Tribrid Mass Spectrometer,[67] and the Orbitrap Astral Mass Spectrometer[68] now enable more comprehensive and high-resolution analyses. These innovations have expanded sample coverage, supporting detailed profiling of diverse post-translational modifications (PTMs), including phosphoproteomics,[69,70] the mapping of subcellular localizations,[71–73] and sampling across a wide variety of cell lines and tissues.[38,44,52,74] Complementing these technological breakthroughs, proteomics analysis software such as MaxQuant,[75] Skyline,[76] Spectronaut,[77] and MSFragger[78]—together with interactive datasets—have streamlined workflows and greatly improved data interpretability in cysteine chemoproteomics.[21,44,74]

While recent advancements in cysteine chemoproteomics have significantly expanded profiling capabilities—detecting up to >78,000 cysteines in a single study—several critical challenges remain. First, the lack of robust pipelines for data processing and integration limits the systematic analysis and cross-comparison of cysteine chemoproteomic datasets. Existing tools often generate separate output folders with numerous files for each biological and technical replicate, relying on comparisons based on protein names, peptide sequences, or peptide-spectrum matches (PSMs), rather than residue-level identifiers, such as Protein_C#. As a result, analyzing whether specific cysteine residues are consistently labeled across replicates often requires time-consuming manual inspection.

Second, there is no rapid or scalable approach to stratify cysteines based on functional importance. As chemoproteomic studies identify thousands of residues, such stratification is crucial for prioritizing targets for follow-up studies. Third, existing approaches lack mechanistic insight into why certain cysteines are predisposed to labeling by electrophilic compounds, a key limitation for facilitating lead development and understanding specific protein-ligand interactions. To address these challenges and advance the identification of functional and druggable cysteines, this dissertation combines mass spectrometry-based cysteine chemoproteomics with the development of novel computational tools.

Among these, MS-CpDAA (Mass Spectrometry-based Chemoproteomics Detected Amino Acids Analysis Suite; https://github.com/lmboat/ms_cpdaa_analysis), is an automation software that expedites identification of covalently targeted residues in high-throughput chemoproteomics experimentation. This tool facilitated the evaluation of a novel experimental approach that combines single-pot, solid-phase-enhanced sample preparation with high-field asymmetric waveform ion mobility spectrometry (SP3-FAIMS) for high-coverage profiling of the human cysteineome.[38] This approach enabled the aggregation of covalent labeling sites across 18 samples, spanning seven cell lines, three proteolytic digestion conditions, and two subcellular fractions, culminating in the detection of 34,225 cysteines. This represents a 5.5-fold increase in coverage compared to prior studies, expanding cysteine coverage from 2% to 13% of the cysteinome.

Beyond profiling advancements, MS-CpDAA has been applied in diverse contexts, including the identification of covalent labeling sites with bifunctional probes in multiplexed CuAAC Suzuki-Miyaura chemoproteomics (mCSCP)[55] and the evaluation of efficiency and selectivity of electrophilic labeling with Tunable Amine-Reactive Electrophiles (TARE probes).[79] The MS-CpDAA output simplifies the integration of datasets, facilitating database consolidation and structured analyses for broader applications.

Building on the ability to aggregate and analyze high-throughput data with MS-CpDAA, Chapter 2 introduces CysDB,[5] a publicly accessible SQL database with an interactive web interface that integrates experimental chemoproteomic measures of cysteine reactivity with protein functional and structural annotations. This resource provides an infrastructure for integrating high throughput chemoproteomic datasets and facilitates the rapid prioritization of functional target proteins.

While CysDB integrates experimental and structural data for target prioritization, Chapter 3 focuses on CIAA (Cysteine reactivity towards IodoAcetamide Alkyne), a framework designed to determine the structural features of proteins promoting elevated cysteine reactivity. This work establishes a foundation for structure-based artificial intelligence (AI) approaches to model protein-ligand interactions, paving the way for future advancements in cysteine-targeted drug discovery.

1.1 - References

1. Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M.F., Kellis, M., Lindblad-Toh, K., and Lander, E.S. (2007). Distinguishing protein-coding and noncoding genes in the human genome. Proc Natl Acad Sci USA *104*, 19428–19433. 10.1073/pnas.0709013104.

2. Park, D., Park, J., Park, S.G., Park, T., and Choi, S.S. (2008). Analysis of human disease genes in the context of gene essentiality. Genomics *92*, 414–418. 10.1016/j.ygeno.2008.08.001.

3. Piovesan, A., Antonaros, F., Vitale, L., Strippoli, P., Pelleri, M.C., and Caracausi, M. (2019). Human protein-coding genes and gene feature statistics in 2019. BMC Res. Notes *12*, 315. 10.1186/s13104-019-4343-8.

4. Amaral, P., Carbonell-Sala, S., De La Vega, F.M., Faial, T., Frankish, A., Gingeras, T., Guigo, R., Harrow, J.L., Hatzigeorgiou, A.G., Johnson, R., et al. (2023). The status of the human gene catalogue. Nature *622*, 41–47. 10.1038/s41586-023-06490-x.

5. 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.

6. Pikarsky, E., Porat, R.M., Stein, I., Abramovitch, R., Amit, S., Kasem, S., Gutkovich-Pyest, E., Urieli-Shoval, S., Galun, E., and Ben-Neriah, Y. (2004). NF-kappaB functions as a tumour promoter in inflammation-associated cancer. Nature *431*, 461–466. 10.1038/nature02924.

7. Porter, A.G., and Jänicke, R.U. (1999). Emerging roles of caspase-3 in apoptosis. Cell Death Differ. *6*, 99–104. 10.1038/sj.cdd.4400476.

8. Arnér, E.S., and Holmgren, A. (2000). Physiological functions of thioredoxin and thioredoxin reductase. Eur. J. Biochem. *267*, 6102–6109. 10.1046/j.1432-1327.2000.01701.x.

9. Amerik, A.Y., and Hochstrasser, M. (2004). Mechanism and function of deubiquitinating enzymes. Biochim. Biophys. Acta *1695*, 189–207. 10.1016/j.bbamcr.2004.10.003.

10. Mahajan, S., Ghosh, S., Sudbeck, E.A., Zheng, Y., Downs, S., Hupke, M., and Uckun, F.M. (1999). Rational design and synthesis of a novel anti-leukemic agent targeting Bruton’s tyrosine kinase (BTK), LFM-A13 [alpha-cyano-beta-hydroxy-beta-methyl-N-(2, 5-dibromophenyl)propenamide]. J. Biol. Chem. *274*, 9587–9599. 10.1074/jbc.274.14.9587.

11. Jorissen, R.N., Walker, F., Pouliot, N., Garrett, T.P.J., Ward, C.W., and Burgess, A.W. (2003). Epidermal growth factor receptor. In The EGF receptor family (Elsevier), pp. 33–55. 10.1016/B978-012160281-9/50004-9.

12. Byrd, J.C., Furman, R.R., Coutre, S.E., Flinn, I.W., Burger, J.A., Blum, K.A., Grant, B., Sharman, J.P., Coleman, M., Wierda, W.G., et al. (2013). Targeting BTK with ibrutinib in relapsed chronic lymphocytic leukemia. N. Engl. J. Med. *369*, 32–42. 10.1056/NEJMoa1215637.

13. Soria, J.-C., Ohe, Y., Vansteenkiste, J., Reungwetwattana, T., Chewaskulyong, B., Lee, K.H., Dechaphunkul, A., Imamura, F., Nogami, N., Kurata, T., et al. (2018). Osimertinib in Untreated EGFR-Mutated Advanced Non-Small-Cell Lung Cancer. N. Engl. J. Med. *378*, 113–125. 10.1056/NEJMoa1713137.

14. Wells, M., Tidow, H., Rutherford, T.J., Markwick, P., Jensen, M.R., Mylonas, E., Svergun, D.I., Blackledge, M., and Fersht, A.R. (2008). Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. Proc Natl Acad Sci USA *105*, 5762–5767. 10.1073/pnas.0801353105.

15. Higueruelo, A.P., Jubb, H., and Blundell, T.L. (2013). Protein-protein interactions as

druggable targets: recent technological advances. Curr. Opin. Pharmacol. *13*, 791–796. 10.1016/j.coph.2013.05.009.

16. Flanagan, S.E., Haapaniemi, E., Russell, M.A., Caswell, R., Allen, H.L., De Franco, E., McDonald, T.J., Rajala, H., Ramelius, A., Barton, J., et al. (2014). Activating germline mutations in STAT3 cause early-onset multi-organ autoimmune disease. Nat. Genet. *46*, 812–814. 10.1038/ng.3040.

17. Boike, L., Cioffi, A.G., Majewski, F.C., Co, J., Henning, N.J., Jones, M.D., Liu, G., McKenna, J.M., Tallarico, J.A., Schirle, M., et al. (2021). Discovery of a Functional Covalent Ligand Targeting an Intrinsically Disordered Cysteine within MYC. Cell Chem. Biol. *28*, 4-13.e17. 10.1016/j.chembiol.2020.09.001.

18. Henley, M.J., and Koehler, A.N. (2021). Advances in targeting "undruggable" transcription factors with small molecules. Nat. Rev. Drug Discov. *20*, 669–688. 10.1038/s41573-021-00199-0.

19. Xie, X., Yu, T., Li, X., Zhang, N., Foster, L.J., Peng, C., Huang, W., and He, G. (2023). Recent advances in targeting the "undruggable" proteins: from drug discovery to clinical trials. Signal Transduct. Target. Ther. *8*, 335. 10.1038/s41392-023-01589-z.

20. Savage, S.R., Yi, X., Lei, J.T., Wen, B., Zhao, H., Liao, Y., Jaehnig, E.J., Somes, L.K., Shafer, P.W., Lee, T.D., et al. (2024). Pan-cancer proteogenomics expands the landscape of therapeutic targets. Cell *187*, 4389-4407.e15. 10.1016/j.cell.2024.05.039.

21. Xiao, H., Jedrychowski, M.P., Schweppe, D.K., Huttlin, E.L., Yu, Q., Heppner, D.E., Li, J., Long, J., Mills, E.L., Szpyt, J., et al. (2020). A Quantitative Tissue-Specific Landscape of Protein Redox Regulation during Aging. Cell *180*, 968-983.e24. 10.1016/j.cell.2020.02.012.

22. Lobell, R.B. (1998). Prenylation of Ras GTPase superfamily proteins and their function in immunobiology. Adv. Immunol. *68*, 145–189. 10.1016/s0065-2776(08)60559-3.

23. 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.

24. Singh, J., Petter, R.C., Baillie, T.A., and Whitty, A. (2011). The resurgence of covalent drugs. Nat. Rev. Drug Discov. *10*, 307–317. 10.1038/nrd3410.

25. 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.

26. 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.

27. 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.

28. Dang, C.V., Reddy, E.P., Shokat, K.M., and Soucek, L. (2017). Drugging the "undruggable" cancer targets. Nat. Rev. Cancer *17*, 502–508. 10.1038/nrc.2017.36.

29. 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. 10.1038/nature12796.

30. Janes, M.R., Zhang, J., Li, L.-S., Hansen, R., Peters, U., Guo, X., Chen, Y., Babbar, A., Firdaus, S.J., Darjania, L., et al. (2018). Targeting KRAS Mutant Cancers with a Covalent G12C-Specific Inhibitor. Cell *172*, 578-589.e17. 10.1016/j.cell.2018.01.006.

31. 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.

32. Bateman, N.W., Goulding, S.P., Shulman, N.J., Gadok, A.K., Szumlinski, K.K., MacCoss, M.J., and Wu, C.C. (2014). Maximizing peptide identification events in proteomic workflows using data-dependent acquisition (DDA). Mol. Cell. Proteomics *13*, 329–338. 10.1074/mcp.M112.026500.

33. 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.

34. 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. 10.1038/nchem.2826.

35. 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.

36. Hughes, C.S., Moggridge, S., Müller, T., Sorensen, P.H., Morin, G.B., and Krijgsveld, J. (2019). Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. Nat. Protoc. *14*, 68–85. 10.1038/s41596-018-0082-x.

37. 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.

38. 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.

39. Johnston, H.E., Yadav, K., Kirkpatrick, J.M., Biggs, G.S., Oxley, D., Kramer, H.B., and Samant, R.S. (2022). Solvent Precipitation SP3 (SP4) Enhances Recovery for Proteomics Sample Preparation without Magnetic Beads. Anal. Chem. *94*, 10320–10328. 10.1021/acs.analchem.1c04200.

40. Xiao, W., Chen, Y., Zhang, J., Guo, Z., Hu, Y., Yang, F., and Wang, C. (2023). A Simplified and Ultrafast Pipeline for Site-Specific Quantitative Chemical Proteomics. J. Proteome Res. *22*, 3360–3367. 10.1021/acs.jproteome.3c00179.

41. Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., Johnstone, R., Mohammed, A.K.A., and Hamon, C. (2003). Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal. Chem. *75*, 1895–1904. 10.1021/ac0262560.

42. Frost, D.C., Feng, Y., and Li, L. (2020). 21-plex DiLeu Isobaric Tags for High-Throughput Quantitative Proteomics. Anal. Chem. *92*, 8228–8234. 10.1021/acs.analchem.0c00473.

43. Li, J., Cai, Z., Bomgarden, R.D., Pike, I., Kuhn, K., Rogers, J.C., Roberts, T.M., Gygi, S.P., and Paulo, J.A. (2021). TMTpro-18plex: The Expanded and Complete Set of TMTpro Reagents for Sample Multiplexing. J. Proteome Res. *20*, 2964–2972. 10.1021/acs.jproteome.1c00168.

44. 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.

45. Burton, N.R., Polasky, D.A., Shikwana, F., Ofori, S., Yan, T., Geiszler, D.J., Veiga Leprevost, F. da, Nesvizhskii, A.I., and Backus, K.M. (2023). Solid-Phase Compatible Silane-Based Cleavable Linker Enables Custom Isobaric Quantitative Chemoproteomics. J. Am. Chem. Soc. *145*, 21303–21318. 10.1021/jacs.3c05797.

46. Mitchell, D.C., Kuljanin, M., Li, J., Van Vranken, J.G., Bulloch, N., Schweppe, D.K., Huttlin, E.L., and Gygi, S.P. (2023). A proteome-wide atlas of drug mechanism of action. Nat. Biotechnol. *41*, 845–857. 10.1038/s41587-022-01539-0.

47. 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.

48. 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.

49. Zuniga, N.R., Frost, D.C., Kuhn, K., Shin, M., Whitehouse, R.L., Wei, T.-Y., He, Y., Dawson, S.L., Pike, I., Bomgarden, R.D., et al. (2024). Achieving a 35-Plex Tandem Mass Tag Reagent Set through Deuterium Incorporation. J. Proteome Res. *23*, 5153–5165. 10.1021/acs.jproteome.4c00668.

50. Ma, T.P., Izrael-Tomasevic, A., Mroue, R., Budayeva, H., Malhotra, S., Raisner, R., Evangelista, M., Rose, C.M., Kirkpatrick, D.S., and Yu, K. (2023). AzidoTMT Enables Direct

Enrichment and Highly Multiplexed Quantitation of Proteome-Wide Functional Residues. J. Proteome Res. *22*, 2218–2231. 10.1021/acs.jproteome.2c00703.

51. Budayeva, H.G., Ma, T.P., Wang, S., Choi, M., and Rose, C.M. (2024). Increasing the Throughput and Reproducibility of Activity-Based Proteome Profiling Studies with Hyperplexing and Intelligent Data Acquisition. J. Proteome Res. *23*, 2934–2947. 10.1021/acs.jproteome.3c00598.

52. 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.

53. McConnell, E.W., Smythers, A.L., and Hicks, L.M. (2020). Maleimide-Based Chemical Proteomics for Quantitative Analysis of Cysteine Reactivity. J. Am. Soc. Mass Spectrom. 10.1021/jasms.0c00116.

54. 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. 10.1021/jacs.9b08831.

55. 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. 10.1021/acs.analchem.0c04726.

56. Huang, F., Han, X., Xiao, X., and Zhou, J. (2022). Covalent warheads targeting cysteine residue: the promising approach in drug development. Molecules *27*.

10.3390/molecules27227728.

57. Liu, S., Yang, Q., Zhang, L., and Luo, S. (2024). Accurate Protein pKa Prediction with Physical Organic Chemistry Guided 3D Protein Representation. J. Chem. Inf. Model. *64*, 4410–4418. 10.1021/acs.jcim.4c00354.

58. Liu, Z., Remsberg, J.R., Li, H., Njomen, E., DeMeester, K.E., Tao, Y., Xia, G., Hayward, R.E., Yoo, M., Nguyen, T., et al. (2024). Proteomic ligandability maps of spirocycle acrylamide stereoprobes identify covalent ERCC3 degraders. J. Am. Chem. Soc. *146*, 10393–10406. 10.1021/jacs.3c13448.

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. 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.

61. Chan, W.C., Liu, X., Magin, R.S., Girardi, N.M., Ficarro, S.B., Hu, W., Tarazona Guzman, M.I., Starnbach, C.A., Felix, A., Adelmant, G., et al. (2023). Accelerating inhibitor discovery for deubiquitinating enzymes. Nat. Commun. *14*, 686. 10.1038/s41467-023-36246-0.

62. 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. 10.1038/s41467-023-39268-w.

63. Röst, H.L., Rosenberger, G., Navarro, P., Gillet, L., Miladinović, S.M., Schubert, O.T., Wolski, W., Collins, B.C., Malmström, J., Malmström, L., et al. (2014). OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nat. Biotechnol. *32*, 219–223. 10.1038/nbt.2841.

64. Chapman, J.D., Goodlett, D.R., and Masselon, C.D. (2014). Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. Mass Spectrom. Rev. *33*, 452–470. 10.1002/mas.21400.

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. 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. 10.1038/s41592-020-00998-0.

67. Shuken, S.R., McAlister, G.C., Barshop, W.D., Canterbury, J.D., Bergen, D., Huang, J., Huguet, R., Paulo, J.A., Zabrouskov, V., Gygi, S.P., et al. (2023). Deep Proteomic Compound Profiling with the Orbitrap Ascend Tribrid Mass Spectrometer Using Tandem Mass Tags and Real-Time Search. Anal. Chem. *95*, 15180–15188. 10.1021/acs.analchem.3c01701.

68. Heil, L.R., Damoc, E., Arrey, T.N., Pashkova, A., Denisov, E., Petzoldt, J., Peterson, A.C., Hsu, C., Searle, B.C., Shulman, N., et al. (2023). Evaluating the Performance of the Astral Mass Analyzer for Quantitative Proteomics Using Data-Independent Acquisition. J. Proteome Res. *22*, 3290–3300. 10.1021/acs.jproteome.3c00357.

69. 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. 10.1038/s41592-022-01398-2.

70. Svenningsen, E., Demir, F., Kromm, F., Rahimic, A., Olagnier, D., Rinschen, M., and Poulsen, T. (2024). Chemoproteomic mapping of the N-terminal cysteinome. 10.26434/chemrxiv-2024-4w549.

71. Bak, D.W., Bechtel, T.J., Falco, J.A., and Weerapana, E. (2019). Cysteine reactivity across the subcellular universe. Curr. Opin. Chem. Biol. *48*, 96–105. 10.1016/j.cbpa.2018.11.002.

72. Yan, T., Julio, A.R., Villanueva, M., Jones, A.E., Ball, A.B., Boatner, L.M., Turmon, A.C., Nguyễn, K.B., Yen, S.L., Desai, H.S., et al. (2023). Proximity-labeling chemoproteomics defines the subcellular cysteinome and inflammation-responsive mitochondrial redoxome. Cell Chem. Biol. *30*, 811-827.e7. 10.1016/j.chembiol.2023.06.008.

73. Yan, T., Boatner, L.M., Cui, L., Tontonoz, P.J., and Backus, K.M. (2023). Defining the Cell Surface Cysteinome Using Two-Step Enrichment Proteomics. JACS Au *3*, 3506–3523. 10.1021/jacsau.3c00707.

74. 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.

75. Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. *26*, 1367–1372. 10.1038/nbt.1511.

76. MacLean, B., Tomazela, D.M., Shulman, N., Chambers, M., Finney, G.L., Frewen, B., Kern,

R., Tabb, D.L., Liebler, D.C., and MacCoss, M.J. (2010). Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics *26*, 966–968. 10.1093/bioinformatics/btq054.

77. Bruderer, R., Bernhardt, O.M., Gandhi, T., Miladinović, S.M., Cheng, L.-Y., Messner, S., Ehrenberger, T., Zanotelli, V., Butscheid, Y., Escher, C., et al. (2015). Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. Mol. Cell. Proteomics *14*, 1400–1410. 10.1074/mcp.M114.044305.

78. 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.

79. Tang, K.-C., Cao, J., Boatner, L.M., Li, L., Farhi, J., Houk, K.N., Spangle, J., Backus, K.M., and Raj, M. (2022). Tunable Amine-Reactive Electrophiles for Selective Profiling of Lysine. Angew. Chem. Int. Ed *61*, e202112107. 10.1002/anie.202112107.

Chapter 2

CysDB: A Human Cysteine Database

Based on Experimental Quantitative Chemoproteomics

Reprinted with permission from

Boatner, L. M.; Palafox, M. F.; Schweppe, D. K.; Backus, K. M.

CysDB: a human cysteine database based on experimental quantitative chemoproteomics.

Cell Chemical Biology **2023**, *30*(6), 683-698.

https://doi.org/10.1016/j.chembiol.2023.04.004. Copyright © 2023 Elsevier Ltd.

Lisa M. Boatner,[1,2] Maria F. Palafox,[3] Devin K. Schweppe,[4] and Keriann M. Backus[1,2,5,6,7,8,9,*]

[1]Biological Chemistry Department, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA, [2]Department of Chemistry and Biochemistry, University of California, Los Angeles, Los Angeles, CA 90095, USA, [3]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA, [4]Department of Genome Sciences, University of Washington, Seattle, WA 98185, USA, [5]Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA 90095, USA, [6]DOE Institute for Genomics and Proteomics, University of California, Los Angeles, Los Angeles, CA 90095, USA, [7]Jonsson Comprehensive Cancer Center, University of California, Los Angeles, Los Angeles, CA 90095, USA, [8]Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, University of California, Los Angeles, Los Angeles, CA 90095, USA, [9]Lead contact *Correspondence: kbackus@mednet.ucla.edu

2.1 - Introduction

Small molecule chemical probes are useful tools for modulating protein function that can serve as leads for future medications. Therefore, ongoing efforts in the chemical biology community have set ambitious goals in matching every protein with a chemical probe.[3] Complicating matters, <4% of the human proteome has been pharmacologically targeted by U.S. Food and Drug Administration (FDA)-approved small molecules. Cysteine chemoproteomics has emerged as an enabling technology that addresses this druggability gap by identifying thousands of functional and potentially druggable cysteines proteome-wide.[1–25] Demonstrating this utility, prior cysteine chemoproteomic studies, including our own, have revealed a strikingly low overlap between proteins containing ‘‘ligandable’’ or potentially ‘‘druggable’’ cysteines and those that have been targeted by FDA-approved molecules.[11]

Cysteine proteomics experiments can be generally classified into four main categories: (1) identification, (2) measuring hyperreactivity, (3) measuring ligandability, and (4) measuring redox state. We consider identification studies as those aiming to increase coverage of cysteine-containing peptides.[4–6] Hyperreactivity experiments measure the intrinsic reactivity of cysteines to ward highly electrophilic probes,[7–10] while ligandability experiments measure the intrinsic ligandability or potential ‘‘druggability’’ of cysteines using libraries of drug-like electrophilic molecules, natural products, and lipid-derived electro philes.[2,11,15–19] Finally, redox protocols are tailored to identify redox-sensitive cysteines.[1,20–23]

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

Despite considerable recent advances in instrumentation, sample preparation, and data analysis, most cysteine chemo proteomics studies only sample a small fraction of all cysteines in the proteome, with the highest coverage studies sampling 13% of all cysteines.[1,7,9] Reasons for this gap include protein abundance and restricted expression profiles, location of cysteines in very long or very short tryptic peptides, which are not detected in standard trypsin digests, and unreactive cysteines, such as those buried within protein cores or located in structural disulfides. Despite these technical limitations, the cysteinome continues to grow, with the addition of multiple high-coverage studies in 2022 alone.[6,10,14]

The availability of easily searchable cysteine databases— including Oximouse,[1] the Ligandable Cysteine Database, and previously reported Cysteinome[24]—has increased the general accessibility of these large proteomics datasets, allowing rapid queries for targets of interest.[9,12,13] However, except for the Cysteinome database, which was launched in 2016 and is no longer publicly accessible, these databases are restricted to datasets derived from single publications.

To facilitate future studies aimed at global or target focused analyses of the cysteinome, we envisioned the establishment of a unified cysteine-focused database that would fulfill the following criteria. First, the database would incorporate datasets from many large-scale cysteinomic studies and therefore enable rapid and facile inter- and intra-dataset comparisons. Second, the database would include information about the reactivity and ligandability of cysteines together with the druggability of their corresponding proteins, as indicated by availability of FDA-approved drugs. Last, and most significant, the database would integrate functional and structural data from UniProtKB/Swiss-Prot, Cancer Gene Census (CGC), ClinVar, Human Protein Atlas (HPA), ChEMBL, DrugBank, and the Protein Data Bank

(PDB),[26–32] to enable prioritization of targets for future studies. Here we present CysDB, which is an interactive database that fulfills these criteria for 62,888 cysteines and 11,621 proteins. Importantly, to promote the continued growth of cysteine chemoproteomics, we also provide a straightforward route for addition of future datasets.

2.2 - Results

2.2.1 - Data curation to establish a set of processed and aggregated chemoproteomics datasets to enable CysDB

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

Notably, all these studies rely on the same general cysteine chemoproteomic workflow: cells or lysates are treated with a cysteine-reactive probe (**Figure 1A**; iodoacetamide alkyne) or an iodoacetamide desthiobiotin reagent (e.g., DBIA[2] or IA-DTB[8]) to cap all accessible cysteines. Labeled proteins are subjected to enrichment on streptavidin or related resins together with sequence-specific proteolysis, followed by liquid chromatography-tandem mass spectrometry. Several of our included studies[7–9] further classify cysteine intrinsic reactivity and pinpoint hyperreactive cysteines by comparing relative cysteine labeling by two concentrations (103 and

13) of a cysteine enrichment handle (**Figures 1A** and **S1**). Signal intensity differences between 100- and 10-mM treated proteomes are reflected by a ratio ($R_{[high]/[low]}$). Hyperreactive cysteines are defined as those with $R_{10:1}$ value <2, indicating labeling events that are not concentration dependent. Most included studies provide a metric of cysteine ligandability or putative druggability,[2,4,5,8,10,11] which is generated by comparing relative labeling by equimolar iodoacetamide in the presence and absence of electrophilic compound, with decreased labeling indicative of a high-occupancy labeling event (**Figures 1A** and **S2**).

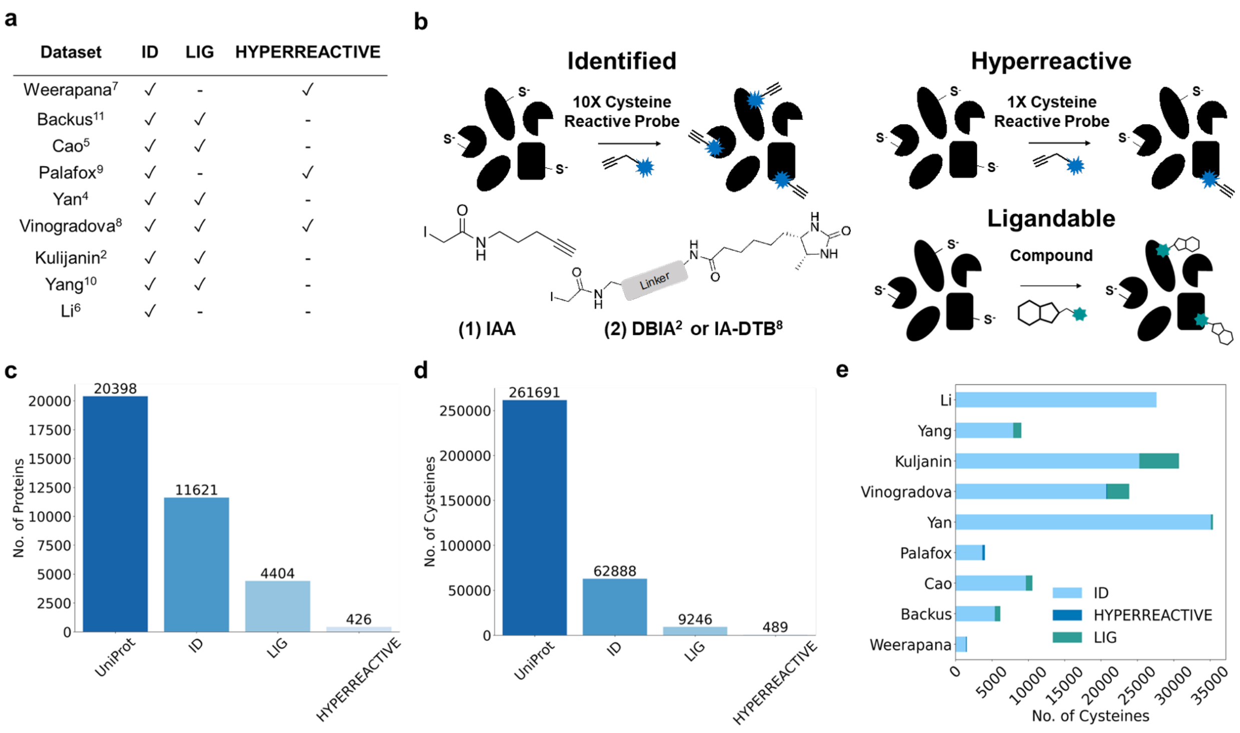
To produce a rigorously curated database, we subjected our prioritized datasets to a series of data-processing steps. First, we aggregated all non-redundant cysteines published by all studies, using the unique identifier UniProtKBID_CYS#. For some studies[2,4–9,11] residue positions and protein identifiers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

Using the studies reporting measures of cysteine ligandability or labeling by electrophilic fragments or drug-like molecules, we further stratified our dataset to generate a master set of all ligandable cysteines. The datasets included in our database (**Figure 1A**) were all prepared using the same general workflow where samples (lysates or cells) were treated by either a vehicle (DMSO) or a cysteine-reactive electrophile functionalized compound and the compound-dependent changes in IAA, DBIA, or IA-DTB reactivity assayed using LC-MS/MS analysis. Prior analyses have revealed that comparable competition ratios can be calculated using either MS1 or MS2 level quantification.[2,4,5,8,10,11] Therefore, we opted not to differentiate between samples analyzed using different quantification methods, including isotopic labeling strategy (TMT or isotopically enriched biotinylation reagents),[2,6] label-free quantification, and data-independent acquisition (DIA) based MS2 level quantification (see **Figure S2** for general workflow).[4,8,10] The vast majority (97.2%) of all compounds screened were found to be functionalized with either a chloroacetamide or acrylamide moieties (**Figure S3**). A small data subset of compounds did, however, feature alternative electrophiles, including covalent reversible cyanoacrylamides,[38] fumarates, and activated esters; although activated esters are primarily lysine reactive, our prior data indicates that they do also exhibit cysteine reactivity.[40,41]

All datasets included in our database relied on competition ratio cutoffs for what defines a cysteine as ‘‘ligandable.’’ Generally, cysteines were categorized as liganded if they had at least two ratios R <= 4 (hit fragments) and one ratio between 0.5 and 2 (control fragments). However, when processing the ligandability data for each dataset, we observed manuscript-specific differences in either the ratio cutoff value or number of minimum unique hit fragments (1 or 2) required to have the associated ratio cutoff value for designating a cysteine as ligandable. For example, Cao et al.[5] implemented a slightly more permissive ratio cutoff of 3 to account for high-field asymmetric waveform ion mobility spectrometry (FAIMS)-induced ratio compression. By comparison, Vinogradova et al.[8] implemented a more stringent ratio cutoff of 5. Another case we encountered was the inclusion of ''ligandable'' cysteines where the unique identifier contained multiple modified cysteine residues, such as UniProtKBID_- CYS#1_CYS#2. These types of identifiers are derived from peptide sequences simultaneously labeled with capture reagents at multiple cysteine residues (C1*XXXC5*) within the same sequence. On the basis of our experience with such peptides yielding noisy ratios, we opted to remove them from CysDB; a total of 2,584 peptides were excluded because of this criterion. Otherwise, despite the differences in defining ligandability, we opted to retain all remaining liganded cysteines to accurately represent each study's reported findings (the criteria for ligand ability applied to each study are available in **Data S1**). In aggregate across all ligandability studies, a total of 43,475 unique cysteines (**Data S2**) had quantified ratios, and 9,246 unique cysteines were deemed ligandable. These cysteines were found in 4,404 proteins (**Figures 1C** and **1D**).

Next, we parsed processed data from published datasets measuring cysteine hyperreactivity.[7–9] The three hyperreactivity studies included in CysDB measured the relative IAA reactivity toward two concentrations of IAA (100 and 10 mM), where a quantitative isoTOP-ABPP ratio ($R_{[high]/[low]}$) reflects the differences in signal intensities between the 100 and 10 mM treated proteomes. Highly reactive cysteines, termed ''hyperreactive'' residues, are identified as those that exhibit saturation or near saturation of labeling at the lower IAA concentration. All three publications used the same numerical ranges to delineate cysteines into ‘‘high,’’ ‘‘medium,’’ and ‘‘low’’ reactivity subsets, with high-reactivity, also termed ‘‘hyperreactive,’’ residues as those with $R_{10:1} < 2$, medium-reactivity cysteines between $R_{100:10}$ R 2 and $R_{10:1} < 5$, and low-reactivity cysteines $R_{10:1} > 5$. During dataset processing, we observed that Weerapana et al.[7] and Pala fox et al.[9] reported median values of all the replicates for each individual measure of cysteine reactivity, as well as an overall mean of medians to quantify the average reactivity per cysteine. In contrast, Vinogradova et al.[8] reported the average of medians across all measurements. To accommodate these dataset dependent differences, we opted to report the mean of median ratio values for each detected cysteine. In aggregate, 8,604 cysteine on 4,032 proteins were quantified by these three studies, which resulted in identification of 489 hyperreactive cysteines and 426 proteins containing hyperreactive cysteines (**Figures 1C** and **1D**).

Collectively across all cysteines identified through our data aggregation efforts, 14.7% were deemed ligandable, and fewer than 1% were determined to be hyperreactive. Cross-dataset comparisons reveal the highest overall coverage dataset was reported by Yan et al.[4] (**Figures 1E** and **S4**), where an optimized SP3-FAIMS strategy was applied to analyze the proteomes of seven cell lines, which in aggregate identified more than 34,000 cysteines on 9,714 proteins from 7 cell lines (**Figures S4** and **S5**). A key outcome of the dataset aggregation required to build CysDB is an effective doubling of the size of the identified cysteinome. Collectively across all studies analyzed in CysDB, ~24% of all cysteines found on 57% of human proteins in UniProtKB have been assayed at least once by chemoproteomics (**Figures 1C** and **1D**).



Dataset	ID	LIG	HYPERREACTIVE
Weerapana[7]	✓	-	✓
Backus[11]	✓	✓	-
Cao[5]	✓	✓	-
Palafox[9]	✓	-	✓
Yan[4]	✓	✓	-
Vinogradova[8]	✓	✓	✓
Kulijanin[2]	✓	✓	-
Yang[10]	✓	✓	-
Li[6]	✓	-	-

Figure 1. Dataset selection and curation for the creation of CysDB. (A) Table of datasets used as input for CysDB, including which datasets were used in each chemoproteomic category (identified, hyperreactive, and ligandable).[2,4–11] (B) General workflows for three categories of chemoproteomic methods included in CysDB that use iodoacetamide alkyne (IAA, **1**) or an iodoacetamide desthiobiotin reagent (DBIA[2] or IA-DTB[8], **2**) to capture cysteines for (1) high-coverage identification of cysteine-containing peptides, (2) quantitative profiling of intrinsic cysteine reactivity, and (3) assaying cysteine ligandability using an electrophile of interest. (C and D) Quantification of the unique proteins (C) and cysteines (D) found in the Human UniProtKB/Swiss-Prot database, together with the identified, ligandable, and hyperreactive chemoproteomics subsets in CysDB. (E) Study-specific breakdown of total number of unique cysteines, including those that are identified as hyperreactive and ligandable. See also **Figures S1–S5** and **Data S1**.

2.2.2 - Establishing an SQL database with an RShiny user interface for CysDB

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

One challenge we faced during our data processing was one-to-one mapping of protein accessions to gene names for SQL querying. For gene-centric queries, not all HUGO Gene Nomenclature Committee (HGNC)[42] or Entrez gene symbols are associated with a single protein. Gene sequences translated to the same protein sequence can lead to multi-mapping of various gene names to one UniProtKB accession.[9] In CysDB, we found that 16 UniProtKB entries were associated with multiple gene names (**Data S1**; **STAR Methods**). To address this limitation, we included the capability to search entries using gene symbols or protein names. The user then selects one of the resulting UniProtKB accessions for CysDB search. The CysDB RShiny interface enables the user to interact with cysteine chemoproteomics datasets, generate personalized figures, and download their results. Anywhere in the app, a user can save graphs as an image by clicking on a camera button at the top right corner and export query results to a CSV (comma-separated value) file by clicking a download button at the bottom of a table. The CysDB app includes five sections: Protein, Mutation, Enrichment, Compound, Statistics, and Datasets.

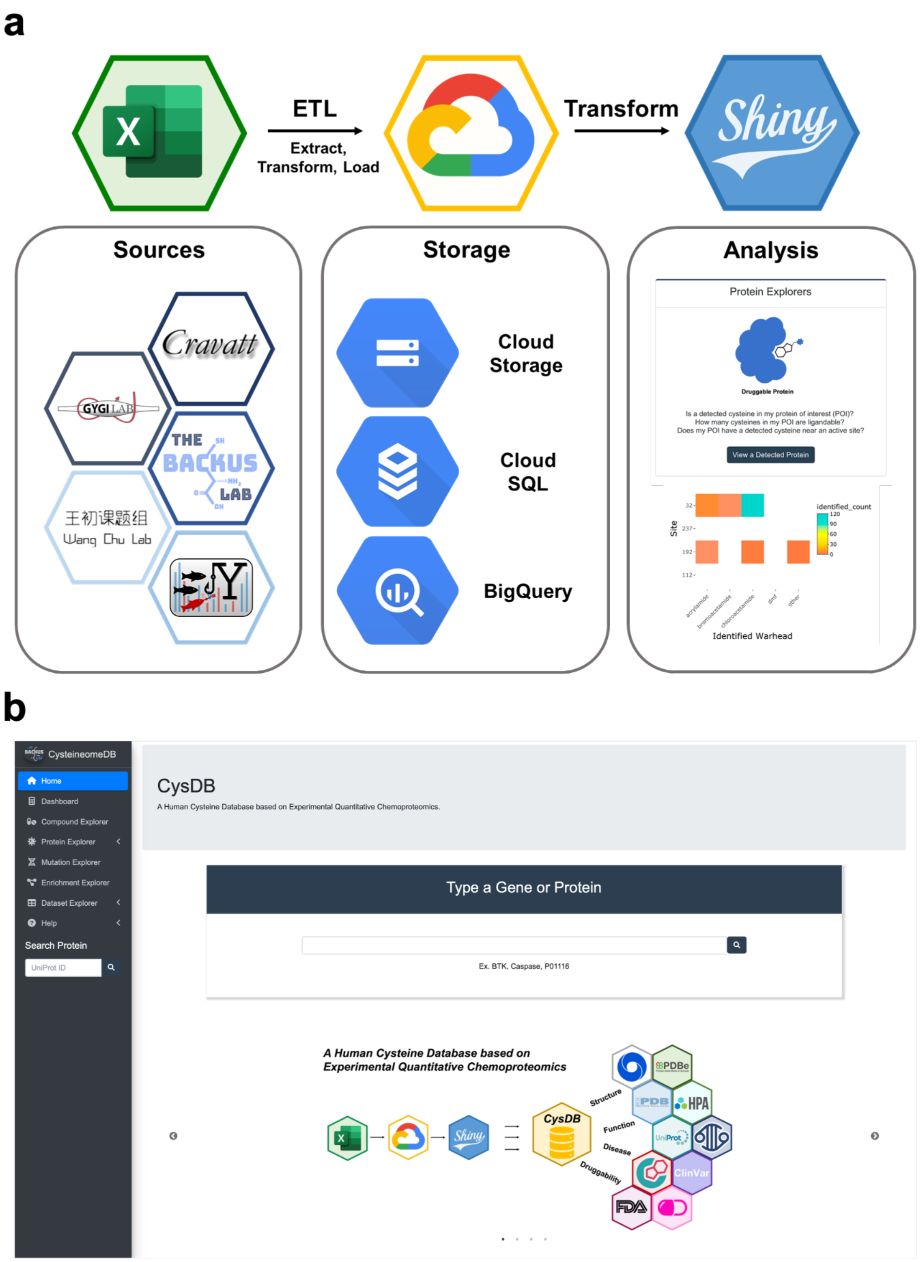
First, users can visualize the CysDB data in a protein-centric manner by selecting the protein explorer button, which is found on the homepage (**Figure 3**). Searching for a protein of interest (POI) by querying a UniProtKB ID returns the ''Protein Section,'' which is further broken up into three separate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

The ''Mutation'' section of CysDB provides information complementary to that presented in the ''Protein Explorer'' section. Querying for a POI yields the aggregate number of CysDB cysteines, missense variants identified in ClinVar,[28] the public repository of relationships between human genetic variation and phenotype, and CGC genes mapped to the POI. This page also generates a one-dimensional depiction of the corresponding protein sequence, decorated with the positions of ligandable and hyperreactive CysDB cysteines alongside individual missense variants, sequence elements, and known ligand binding sites (**Figure 4A**). To further enable pinpointing of cysteines relevant to human health, CysDB also provides CGC annotations of tumor types associated with POI, where relevant.

Looking beyond individual POIs, the ‘‘Enrichment’’ section of CysDB was built for facile visualization and analysis of aggregated ligandable and hyperreactive CysDB subsets. Global analyses provided powered by the Enrichr package include functional pathways, ontologies, and disease enrichments of CysDB categories (**Figure 4B**).[45,46] As with the dataset-wide meta-analysis provided by the Enrichment section, the ‘‘Compound’’ section of CysDB provides users with a global perspective of the electrophilic compounds employed in the CysDB cysteine ligandability studies. This portion of CysDB includes details of each molecule used in the ligandability experiments, including the publication name of each compound, corresponding CysDB names for each corresponding compound and data set in an easily downloadable table.

For the ‘‘Compound’’ section, results can be searched on the basis of SMILES strings or newly created identifiers, defined by a unique combination of SMILES strings, cell lines, and publication authors. Consistent with previous studies,[47] we found that the molecular connectivity for a single two-dimensional (2D) chemical structure could be written in various forms (e.g., ethanol can be denoted as C(O)C, as well as CCO). Thus, we transformed the SMILES strings extracted from each publication into 2D chemical structures and converted these 2D chemical structures into new SMILES strings using RDKit. Selection of a compound identifier using the provided drop down menus, affords a two-dimensional rendering of the chemical structure and computed properties of ‘‘drug-likeness,’’ including the number of hydrogen bond donors and acceptors (**Figure 4C**).[48–53] For this section, we created two separate CysDB compound identifiers to produce scatter plots showing the highest ratios collected for each compound.

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

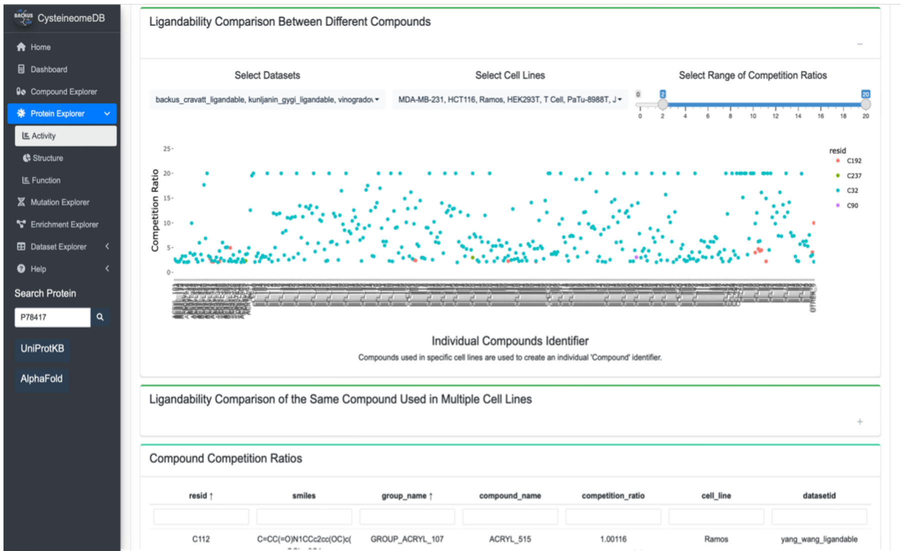


Is a detected cysteine in my protein of interest (POI)?
How many cysteines in my POI are ligandable?
Does my POI have a detected cysteine near an active site?

A Human Cysteine Database based on Experimental Quantitative Chemoproteomics.

A Human Cysteine Database based on Experimental Quantitative Chemoproteomics

Figure 2. Workflow to generate CysDB SQL database. (A) Data extracted from nine datasets (**Data S1**) was transformed and loaded into a MySQL relational database on the Google Cloud Platform. An accompanying front-end web interface was developed using RShiny to allow remote-user querying of the SQL database. (B) Homepage of the CysDB app publicly available at https://backuslab.shinyapps.io/cysdb/. See **Figure S6** and **Data S1** and **S2**.

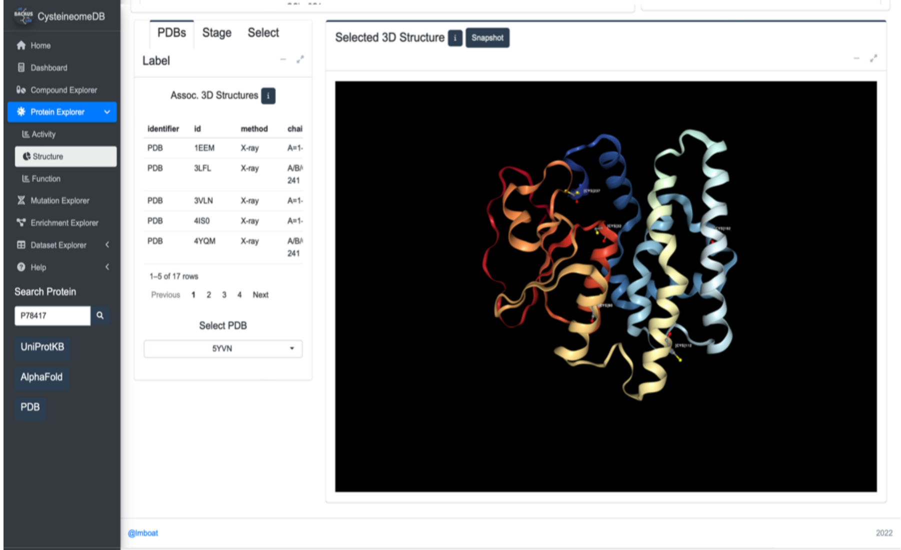

CysteineomeDB
Home
Dashboard
Compound Explorer
Protein Explorer
Activity
Structure
Function
Mutation Explorer
Enrichment Explorer
Dataset Explorer
Help
Search Protein
P78417
UniProtKB
AlphaFold
Select Datasets
backus_cravatt_ligandable, kunljanin_gygi_ligandable, vinogradov
Select Cell Lines
MDA-MB-231, HCT116, Ramos, HEK293T, T Cell, PaTu-8988T, J
Select Range of Competition Ratios
0
2
20
0
2
4
6
8
10
12
14
16
18
20
Competition Ratio
25
20
15
10
5
0
resid
C192
C237
C32
C90
Individual Compounds Identifier
Compounds used in specific cell lines are used to create an individual 'Compound' identifier.

Ligandability Comparison Between Different Compounds

Ligandability Comparison of the Same Compound Used in Multiple Cell Lines

Compound Competition Ratios

resid ↑	smiles	group_name ↑	compound_name	competition_ratio	cell_line	datasetid
C112	C=CC(=O)N1CCc2cc(OC)c(	GROUP_ACRYL_107	ACRYL_515	1.00116	Ramos	yang_wang_ligandable



Assoc. 3D Structures

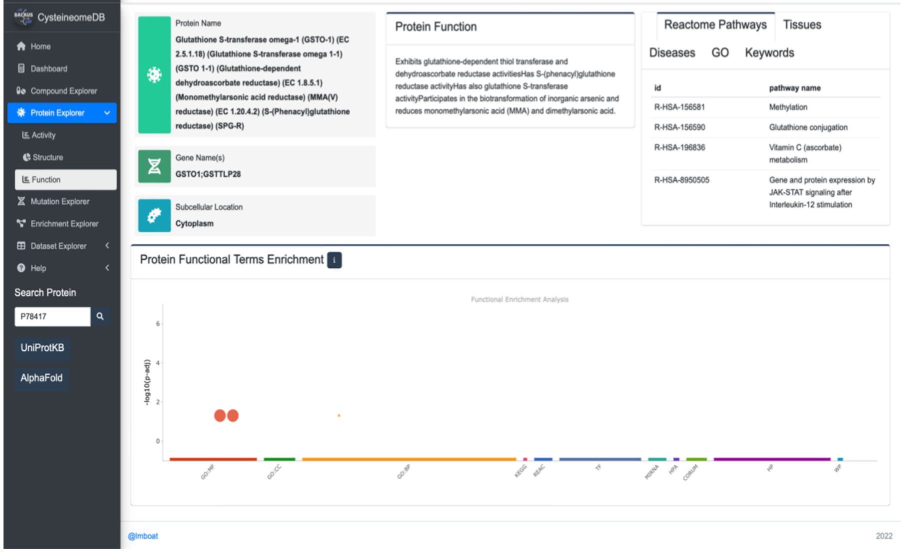
identifier	id	method	chai
PDB	1EEM	X-ray	A=1-
PDB	3LFL	X-ray	A/B/ 241
PDB	3VLN	X-ray	A=1-
PDB	4IS0	X-ray	A=1-
PDB	4YQM	X-ray	A/B/ 241

1–5 of 17 rows

Select PDB: 5YVN

Selected 3D Structure

@lmboat 2022



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

Gene Name(s): **GSTO1;GSTTLP28**

Subcellular Location: **Cytoplasm**

Protein Function

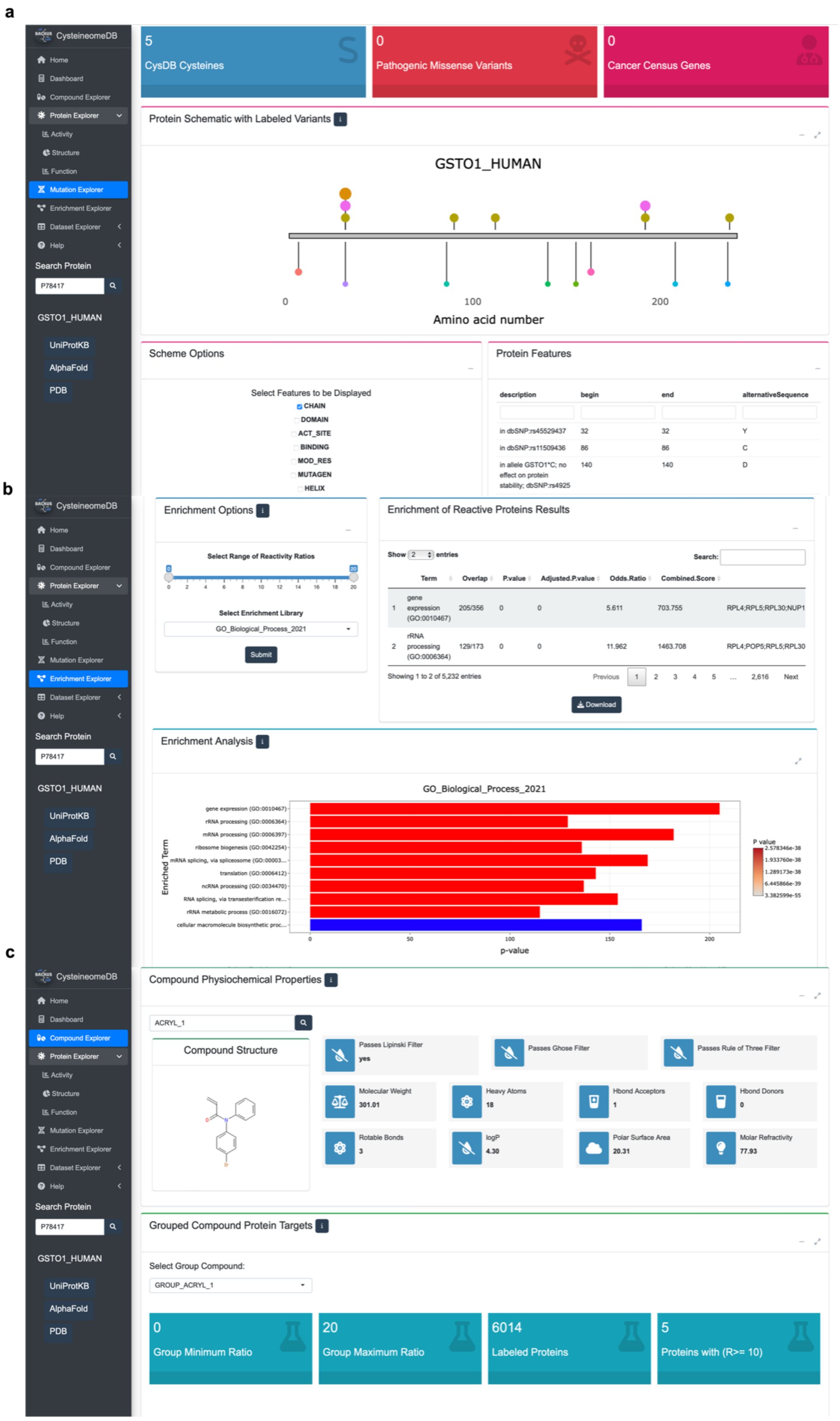
Exhibits glutathione-dependent thiol transferase and dehydroascorbate reductase activitiesHas S-(phenacyl)glutathione reductase activityHas also glutathione S-transferase activityParticipates in the biotransformation of inorganic arsenic and reduces monomethylarsonic acid (MMA) and dimethylarsonic acid.

id	pathway name
R-HSA-156581	Methylation
R-HSA-156590	Glutathione conjugation
R-HSA-196836	Vitamin C (ascorbate) metabolism
R-HSA-8950505	Gene and protein expression by JAK-STAT signaling after Interleukin-12 stimulation

Protein Functional Terms Enrichment

@lmboat 2022

Figure 3. CysDB enables protein-centric queries. Users can search for a protein of interest (POI) in the search bar on the protein page. Centered on the activity tab is a ‘‘site map,’’ indicating which cysteines have been identified, liganded, or hyperreactive by chemoproteomics. In addition, the activity tab allows users to assess the potential druggability of their POIs through scatter plots and heatmaps for quantitative chemoproteomic measurements. For a comprehensive view of the structural environment surrounding the chemoproteomic detected cysteines, publicly available 3D crystal structures are displayed in the structure tab. Users can choose which structure is shown and add customized labels. By clicking the function tab, one can view general information on the POI, including subcellular locations, functional pathways, and GO/Kyoto Encyclopedia of Genes and Genomes (KEGG) terms.



a

Search Protein: P78417

GSTO1_HUMAN

5 CysDB Cysteines

0 Pathogenic Missense Variants

0 Cancer Census Genes

Protein Schematic with Labeled Variants

Scheme Options

Select Features to be Displayed

☑ CHAIN
□ DOMAIN
□ ACT_SITE
□ BINDING
□ MOD_RES
□ MUTAGEN
□ HELIX

Protein Features

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

b

Search Protein: P78417

GSTO1_HUMAN

Enrichment Options

Select Range of Reactivity Ratios: 0 – 20

Select Enrichment Library: GO_Biological_Process_2021

Submit

Enrichment of Reactive Proteins Results

Show 2 entries

Search:

	Term	Overlap	P.value	Adjusted.P.value	Odds.Ratio	Combined.Score	
1	gene expression (GO:0010467)	205/356	0	0	5.611	703.755	RPL4;RPL5;RPL30;NUP1
2	rRNA processing (GO:0006364)	129/173	0	0	11.962	1463.708	RPL4;POP5;RPL5;RPL30

Showing 1 to 2 of 5,232 entries

Previous 1 2 3 4 5 … 2,616 Next

Download

Enrichment Analysis

c

Search Protein: P78417

GSTO1_HUMAN

Compound Physiochemical Properties

ACRYL_1

Passes Lipinski Filter: yes

Passes Ghose Filter

Passes Rule of Three Filter

Molecular Weight: 301.01

Heavy Atoms: 18

Hbond Acceptors: 1

Hbond Donors: 0

Rotable Bonds: 3

logP: 4.30

Polar Surface Area: 20.31

Molar Refractivity: 77.93

Grouped Compound Protein Targets

Select Group Compound: GROUP_ACRYL_1

0 Group Minimum Ratio

20 Group Maximum Ratio

6014 Labeled Proteins

5 Proteins with (R>= 10)

Figure 4. CysDB enables disease, dataset, and cysteine-reactive compound wise queries. (A) The disease relevance of a POI can be explored through the mutation page. Proximity of chemoproteomic detected cysteines, annotated small molecule binders and variants of ranging clinical significance are visualized on a one-dimensional schematic of a protein sequence. Chemoproteomic cysteines are colored in gold for identified, pink for ligandable and orange for hyperreactive, while the remaining points are variant positions. (B) Users can specify subsets of data available in CysDB, such as by compound chemotype or ranges of reactivity ratio, for pathway, ontology, and disease enrichment analyses. The results can then be downloaded as a CSV-formatted table or a bar graph as an image. (C) Chemical structures and calculated ‘‘drug-likeness’’ properties of compounds used to ligand cysteines in CysDB can be accessed from the dropdown menu in the compound page.

2.2.3 - Understanding the scope of the CysDB ligandable or putative ‘‘druggable’’ proteome

We further parsed the data available in CysDB to showcase features built into CysDB and to facilitate the identification of new potential targets for future chemical probe development campaigns. More broadly, we also seek to highlight future opportunities for the cysteine chemoproteomic community. Given the low overlap between FDA-approved drug targets and proteins labeled by cysteine-reactive compounds for prior smaller cysteine chemoproteomics studies,[11] we next extended this analysis to CysDB. Fewer than 4% of all human proteins in UniProtKB have been targeted by FDA-approved small molecules (**Figure S7**). As only 14.7% of all cysteines in CysDB were reported as likely ligandable, we next performed the same analysis on the subset of proteins in CysDB that contain a ligandable cysteine. Again, consistent with the prior reports that have demonstrated a low overlap between targets of covalent compounds and FDA-approved drugs, we find that 3% of proteins that contain one or more ligandable cysteine have been targeted by FDA-approved drugs (**Figure 5A**). Broadening this analysis to a less restrictive set of compound-protein interactions, we find that 32.5% of proteins with ligandable cysteines have been targeted by small molecules, as reported by ChEMBL, DrugBank, and the FDA (**Figure 5B**). These findings showcase the opportunities for targeting undrugged proteins using cysteine-reactive chemical probes.

Prior studies have shown that drug and putative drug targets are highly enriched for protein classes featuring well-defined binding sites, including enzymes and receptors. Therefore, we characterized whether the CysDB members represent new drug gable space by parsing UniProtKB keyword functional annotations of all ligandable proteins in CysDB. Stratification of the CysDB ligandable proteins into two categories, targeted and untargeted by FDA-approved compounds, acknowledged an enrichment for enzymes in the FDA-approved subset (**Figure 5C**). In contrast, the functions of the non-FDA subset of ligandable proteins in CysDB span important protein classes, including transcription factors (TFs), which are often categorized as a largely ‘‘undruggable’’ class of proteins, with the notable exception of TFs with well-defined small molecule ligand binding pockets, such as nuclear hormone receptors.

Next, we analyzed the compounds that target ligandable cysteine residues to further dissect the potential druggability of CysDB entries. Several different electrophilic moieties, often termed ‘‘warheads,’’ have been developed, which react with cysteine residues in both irreversible and covalent reversible modes of labeling.[39,54–56] Examples of these electrophilic handles include compounds that react via a thiol-Michael addition (e.g., irreversible modifiers such as acrylamide, fumarate esters, vinyl sulfonamide together with reversible modifiers such as cyanoacrylamide), compounds that react via S_N2 (e.g., alpha-halo compounds), as well as compounds that react via S_NAr (e.g., halogen-substituted electron deficient heterocycles such as chlorotriazine). As prior studies have revealed varying proteome-wide reactivity and structure-activity relationships (SARs) for different cysteine-reactive electrophiles, we decided to quantify the number of cysteines detected as labeled by individual electrophile chemotypes. For this analysis, a cysteine was labeled by one of the five warheads if the cysteine had $R <= 4$ for at least one compound (**Figures 5D**, **S8**, and **S9**).[2,27–62] Over all, we found that a large majority of the ligandability data were acquired for samples subjected to labeling by acrylamides (AAs) and chloroacetamide (CA)-substituted compounds across the panel of cell lines tested (**Figures 5D** and **S10**), with a small fraction derived from additional probes ranging from cyanoacrylamides to dimethyl fumarate listed in **Data S2**. Interestingly, we noticed that some cysteines react promiscuously with both AA and CA electrophiles, whereas others show an electrophile preference (**Figure 5E**). The proteins glutathione S-transferase omega-1 (GSTO1) and carbonyl reductase (CBR1) exemplify the striking electrophile preference observed for some proteins (**Figure 5F**). For GSTO1, the highly ligandable cysteine (Cys 32) exhibits strong preference for reacting with chloroacetamide substituted compounds (1 to 11.5 in favor of CA electrophiles, with respect to unique SMILES strings with the CA moiety). In contrast, cysteine 226 of CBR1 shows marked acrylamide bias (5 to 1 in preference of AA warheads, with respect to unique SMILES strings with the AA moiety).

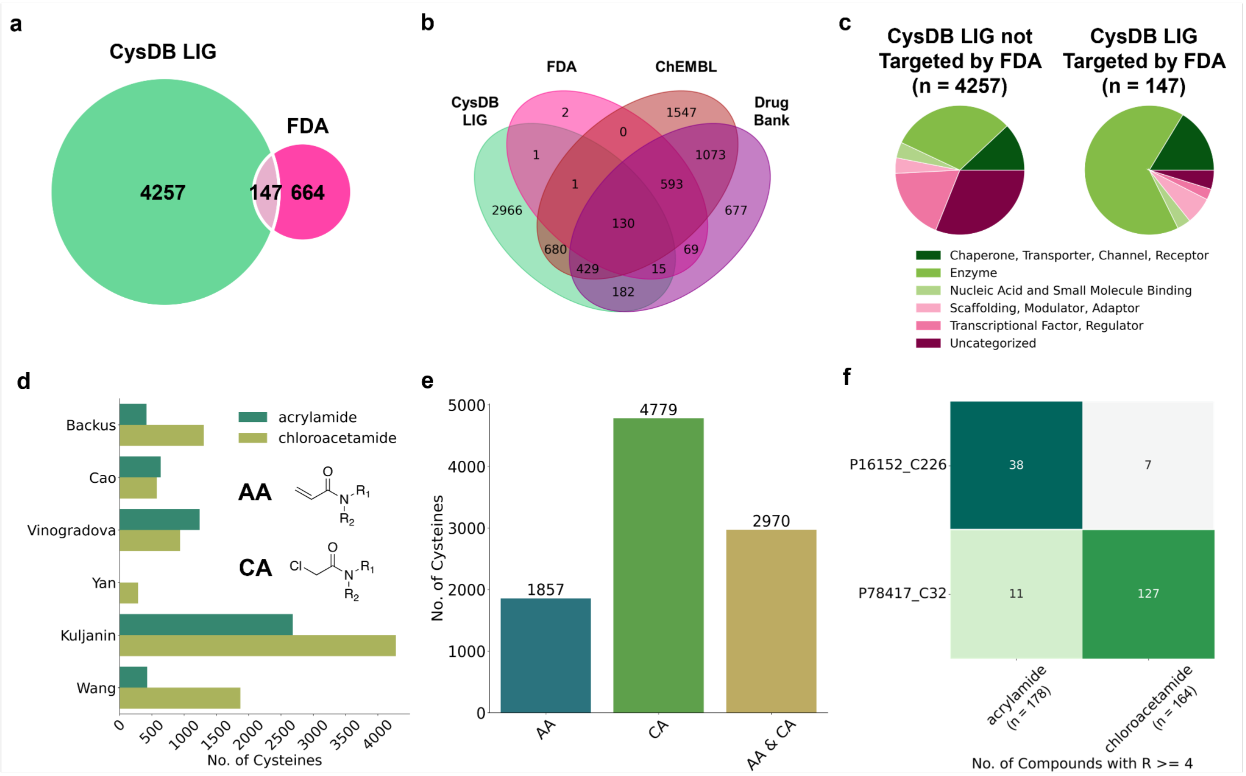


Figure 5. Cysteines with available ligandability data. (A) Overlap between CysDB ligandable (LIG) proteins and proteins targeted by FDA-approved drugs. (B) Overlap between CysDB LIG proteins, proteins targeted by FDA-approved drugs, small molecules in DrugBank and ChEMBL. (C) Distributions of protein functions for CysDB LIG proteins not targeted by FDA and CysDB LIG proteins targeted by FDA. (D) Grouped bar graph showing the number of unique ligandable cysteines targeted by acrylamides or chloroacetamide for each dataset (R <= 4 for at least one compound). (E) Bar graph of the overall number of unique cysteines targeted by acrylamides or chloroacetamide. (F) Number of unique SMILES strings with an acrylamide and chloroacetamide moiety (on the basis of the ‘‘group compound identifier’’), compounds with ratios R <= 4 for protein carbonyl reductase (CBR1; UniProtKB: P16512) and protein glutathione s-transferase omega-1 (GSTO1; UniProtKB: P78417). See also **Figures S7–S10** and **Data S2**.

2.2.4 - Characterizing CysDB proteins on the basis of structural, activity, and functional annotations

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

We first harnessed UniProtKB annotations to determine which CysDB proteins had functional annotations of the following active sites, binding sites, catalytic activity, disulfide bonds, and redox potentials. Analysis concluded 1,505 CysDB proteins possess an active site, 2,961 possess a binding site, 2,784 have experimental evidence for catalytic activity, 1,077 have annotated disulfide bonds and 52 have experimental evidence for redox potentials (**Figure 6A**). Comparable distribution of functional annotations was observed when stratifying the CysDB dataset to consider hyperreactive and ligandable proteins.

To assess whether any CysDB cysteines were annotated as known active or binding sites, we parsed the UniProtKB site and notations for residue positions. This analysis uncovered that, while cysteine is a relatively rare amino acid (2.3% of all proteinaceous amino acids are cysteines[1]), cysteine is the second most abundant binding site amino acid and the third most abundant active site amino acid (**Figures S11** and **S12**). Overall, CysDB reports identification of

1,335 (31.8%) of all known cysteine matching UniProtKB annotated binding sites and 288 (49%) of all known cysteine-active sites (**Figure 6B**). Of the 4,198 cysteine specific binding sites, 178 of them have been liganded by a compound in CysDB. In addition, 98 out of the 583 cysteine-active sites have been liganded by a compound in CysDB and 41 out of the 583 cysteine-active sites were deemed hyperreactive (**Figure S13**).

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

As the UniProtKB dataset is limited to 1D analysis, we asked whether CysDB could also provide insight into the three-dimensional (3D) microenvironment of identified cysteines, using structures reported in the PDB. In total, 5,270 CysDB ID proteins are associated with an available PDB structure, which represents 70% of all human genes with available crystallographic structures (**Figure S16**). Of these, 2,314 (31%) contain one or more ligandable cysteines and 279 feature at least one hyperreactive cysteine (**Figure 6C**). To confirm whether a CysDB cysteine was resolved in a PDB structure, we parsed the residue numbers and coordinates from PDB files. To account for discrepancies between UniProtKB and PDB residue numbers, residue to protein sequence numbering was mapped using SIFT annotations[63] (**Figure S16**). This systematic analysis of residue-level mapping established that out of all the proteins with annotated binding or active sites, 2,684 and 1,315 proteins, respectively, are associated with PDB structures (**Figures S17** and **S18**; **STAR Methods**). Of these, 1,007 proteins have cysteine-binding sites resolved in a corresponding structure, while 338 proteins have cysteine-active sites resolved in a corresponding structure. In aggregate, 18,959 (30.1%) of CysDB-identified cysteines are resolved in a corresponding crystal structure. Further inspection of this dataset revealed that 1,212 CysDB cysteines are proximal (within 10 Å) to binding site residues and 704 CysDB cysteines are proximal to active site residues in 3D space (**Figures S19** and **S20**; **STAR Methods**). To assist structure-guided analysis of cysteine datasets, CysDB provides users with 3D interactive renderings of cysteine-containing structures that include known functional annotations.

Notably, 8,214 proteins (71%) identified by chemoproteomics do not have highly supported evidence in UniProtKB for binding or active sites. Therefore, we next asked whether the CysDB platform could provide additional information about these proteins and corresponding identified cysteines to further aid in delineation of functionally significant cysteines. To guide our platform development efforts, we tested whether the ligandable and hyperreactive cysteine-containing protein subsets are enriched for particular structural domains and functional pathways. Enrichment analysis of protein family (Pfam)[64] domains elucidated a 13-fold enrichment of liganded proteins in the DEAD/DEAH box helicase family, which is consistent with our prior observation of enrichment for RNA binding proteins in chemoproteomics datasets (**Figure 6D**).[65] Responsible for unwinding the duplex of double-stranded RNA, mutations in DEAD/DEAH proteins have been linked to autoimmune disease and some cancers, such as DEAD-box helicase 3 X-linked (DDX3X) in medulloblastoma.[66–69] Pfam domain enrichment analysis for the hyperreactive cysteine subset, revealed an enrichment of thioredoxin and arginine kinase families. These findings are consistent with prior reports of redox enzymes featuring highly reactive cysteines.[7] Notably, creatine kinase enzymes are members of the arginine kinase family of enzymes, which are known to have highly reactive active site cysteines.[7]

We then extended these studies to Panther[70] pathway analysis to assess if pathways are enriched for reactive or ligandable cysteines. We observe an enrichment of ligandable cysteine containing proteins implicated in apoptosis (**Figure 6E**). Examples of ligandable cysteine-containing proteins include TP53, caspase-8, and APBB2. Given the central relevance in modulating cell death to treat numerous disorders, including cancers and neurodegenerative disorders, we expect that this observed notable enrichment indicates untapped opportunities for the development of probes targeting cell death.[71,72] The hyperreactive cysteine-containing protein set, by contrast, was distinctly enriched for proteins involved in integrin signaling. These findings are consistent with the enrichment for hyperreactive cysteines in thioredoxin proteins and related antioxidant systems that are critical for regulation of integrin abundance, secretion, and disulfide formation.[73,74]

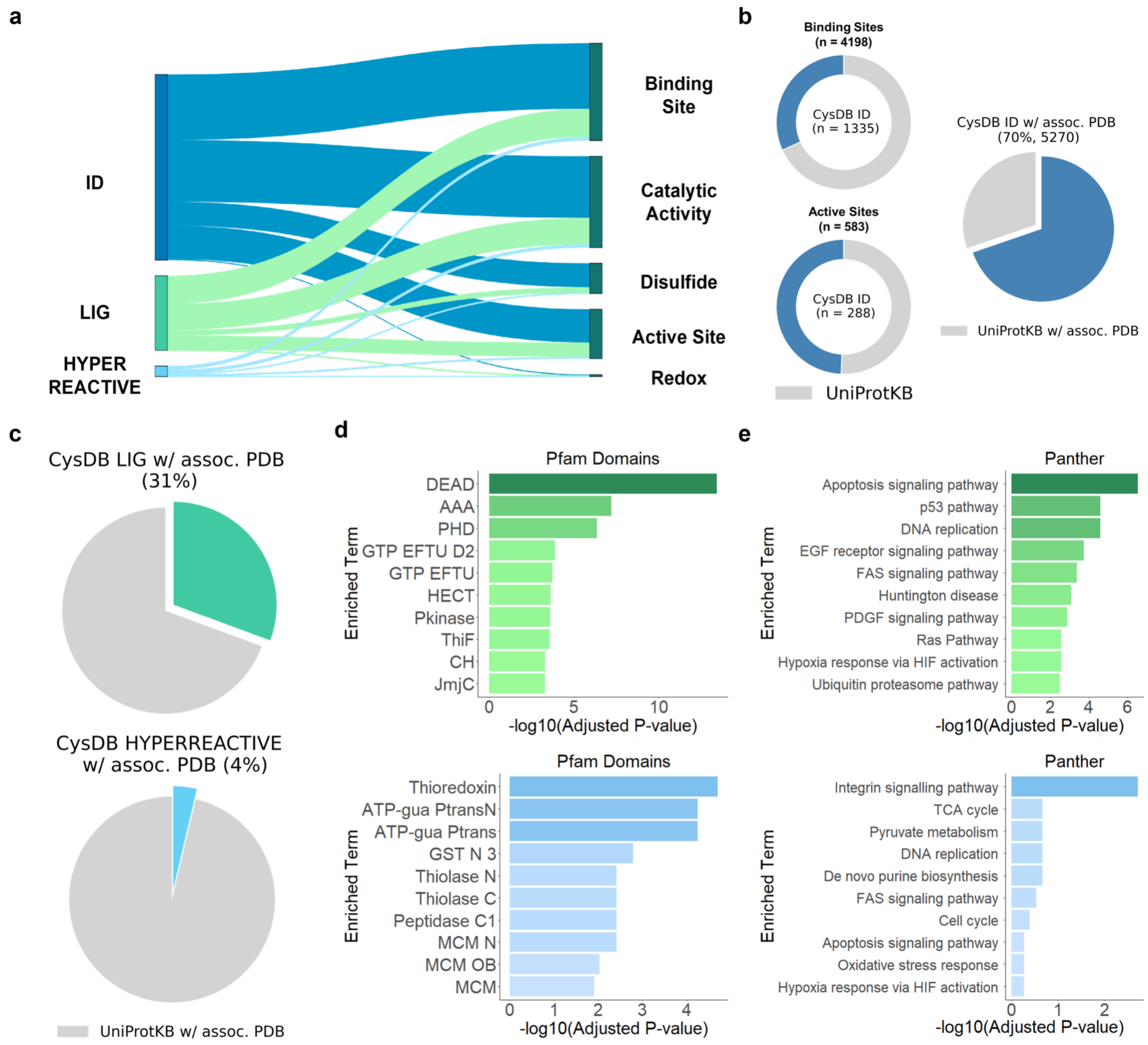


Figure 6. Cysteines with available functional and structural annotations. (A) CysDB-identified, ligandable, and hyperreactive proteins with annotated active sites, binding sites, catalytic activity, disulfide bonds, and redox potentials. (B) Distribution of identified cysteines in CysDB ID annotated as cysteine-specific binding sites or active sites (left). The total number of cysteines in UniProtKB annotated as binding or active sites are shown in gray. Percentage of proteins associated with a PDB structure and containing an identified cysteine. (C) Percentage of proteins associated with a PDB structure and containing a ligandable (CysDB LIG) or

hyperreactive (CysDB HYPERREACTIVE) cysteine. (D) Top 10 enriched protein domains from Pfam term enrichment analysis of liganded (green) and hyperreactive (light blue) proteins. (E) Top 10 enriched pathways from Panther term enrichment analysis of liganded (green) and hyperreactive (light blue) proteins. See also **Figures S11–S20** and **Data S3**.

2.2.5 - Stratifying CysDB proteins on the basis of disease relevant annotations, including cancer association and measures of genetic variation

Building upon our analyses of protein function, we assessed the human disease relevance of the CysDB proteins. Restricting our analysis to the ligandable and hyperreactive subsets, we analyzed which phenotypes were associated with CysDB proteins. Using disease annotations from the Online Mendelian Inheritance in Man (OMIM)[75] knowledge base, ligandable cysteine-containing proteins showed terms related to a broad range of cancers, including colorectal, breast, and leukemia. The hyperreactive cysteine-containing protein subset was enriched for terms associated with immune-relevant diseases, specifically those affecting the lymphatic system (**Figure S25**). Next, we determined how many CysDB proteins are annotated as cancer-driving genes, as dictated by the CGC.[27] Seventy-six percent of CGC genes have been identified by CysDB (559/733) (**Figure S28**). Of all the CGC genes, 38% are annotated as ligandable in CysDB, indicating untapped opportunities for the development of tailored therapies targeting driver mutations (**Figure 7A**; **Data S4**). These results compare favorably with the 11% of cancer-driving genes that have been targeted by FDA-approved small molecules (**Figure S29**; **Data S2**). We observed a considerable difference in the number of available therapies for different cancers during our enrichment analysis for CysDB proteins associated with different tumor types. Although acute myeloid leukemia (AML) genes are the most represented somatic tumor type in CGC, only 5% of these genes are targets of FDA-approved small molecules. By contrast, 13 out of 38 (34%) of non-small cell lung cancer (NSLC) genes have been targeted by FDA-approved drugs. Toward addressing this therapy gap, CysDB detects most CGC genes associated with AML, 71 out of 81 (88%) (**Figure 7B**). In fact, 36 of these AML genes have been liganded by a compound in CysDB, such as the class 2 AML genes nucleophosmin 1 (NPM1) and core-binding factor subunit beta (CBFB).

Genetic variants, along with wild-type genes, can contribute toward harmful disease phenotypes. The ClinVar[28] database provides a curated set of clinical significance for more than 1 million genetic variants, which are classified as benign, pathogenic, or variants of unknown significance (VUS). Of 12,858 unique UniProtKB proteins associated with ClinVar variants (mapped to 31,685 unique genes), 9,478 proteins (73.7%) have a missense variant (**Figure S30**). Overall, more than half of the proteins identified in CysDB have an associated ClinVar missense variant, of which 3,075 contain liganded cysteines and 330 contain hyperreactive cysteines (**Figure 7C**). Previously we reported a trend between chemoproteomic identified cysteines and missense pathogenicity, where chemoproteomic detected cysteine codons were predicted to be more deleterious than undetected cysteine codons.[9] Consistent with the ubiquity of missense variants in ClinVar, the most common mutation associated with CysDB ID CGC genes are missense mutations.[27] Of the CysDB ID proteins that have a ClinVar missense variant, 4,418 proteins have a benign variant, 2,524 proteins have a pathogenic variant, and 3,333 proteins have a variant of unknown significance (**Figure S31**). The proteins with the highest number of pathogenic variants are fibrillin-1 (FBN1; UniProtKB: P35555) and low-density lipoprotein receptor (LDLRl UniProtKB: P01130) (**Figure 7D**). Mutations in FBN1 are known to frequently cause Marfan syndrome by destabilizing disulfide bonds of conserved cysteine residues in epidermal growth factor (EGF)-like domains.[76–78] Additionally, LDLR contains cysteine-rich repeats that bind lipoproteins. Loss-of-function mutations in these regions result in the disruption of cholesterol transport, leading to an increased risk for heart disease.[79,80] In addition to enabling human genotype-guided target prioritization, targeting variant containing chemoproteomic detected proteins may also prove useful precision therapy development in a manner akin to the recent Gly12Cys-directed KRAS compounds, including FDA approved Sotorasib.[81–83]

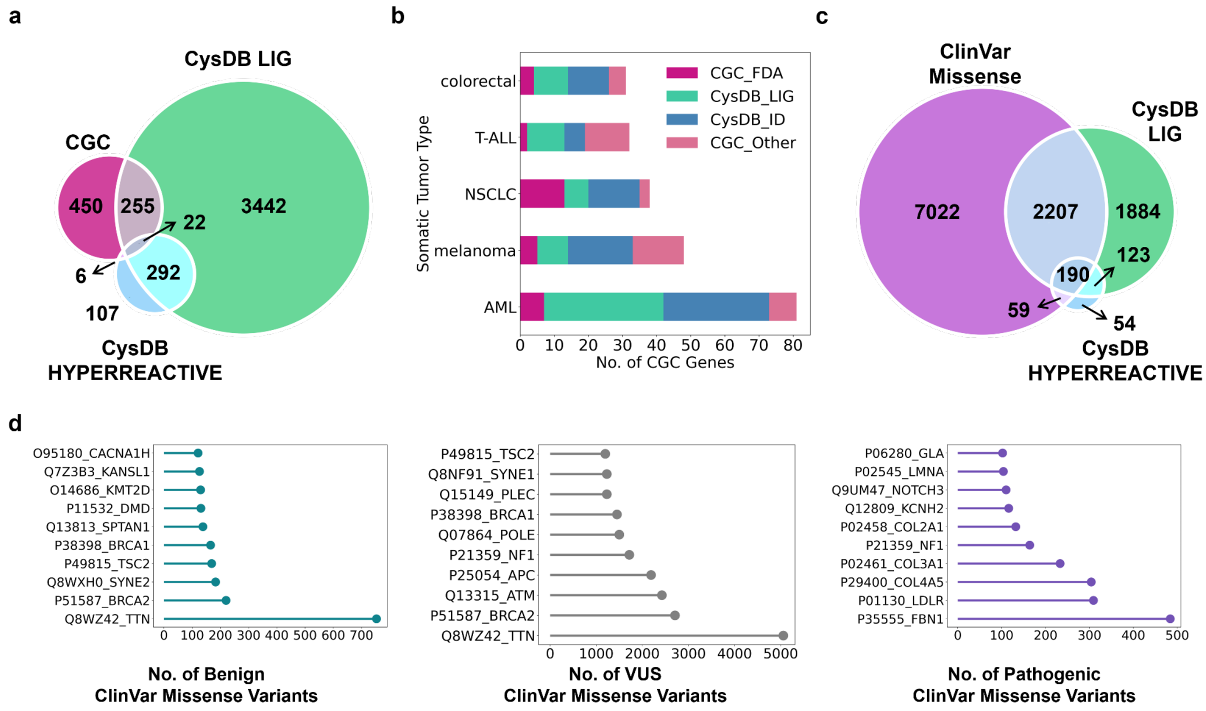


Figure 7. Assessment of the scope of disease-relevant proteins contained in CysDB of biologically relevant proteins using cysteine chemoproteomics. (A) Overlap between genes associated with cancer by the Cancer Gene Census (CGC), genes associated with CysDB ligandable proteins, and genes associated with CysDB hyperreactive proteins. (B) For the five most abundant tumor types in CGC, the number of CGC genes targeted by FDA-approved drugs (CGC_FDA), non-FDA targeted CGC genes identified in CysDB (CysDB_ID), non-FDA targeted CGC genes liganded in CysDB (CysDB_LIG), and non-FDA targeted CGC genes not identified in CysDB (CGC_Other). (C) Overlap between unique proteins associated with ClinVar genes containing missense variants (9,951 genes mapped to 9,478 proteins), CysDB ligandable proteins, and CysDB hyperreactive proteins. (D) Top ten CysDB identified proteins with the highest number of benign missense variants (teal), missense variants of unknown significance (VUS) (gray), and pathogenic missense variants (purple). See also

Figures S21–S31 and **Data S4**.

2.3 - Discussion

Leading groups in cysteine chemoproteomics have discovered thousands of functional and potentially druggable cysteines proteome-wide.[1–9] These studies have yielded global measures of the SAR of compounds that target specific cysteines together with the intrinsic reactivity toward promiscuous electrophilic probes. Given the functional and clinical significance of identification of reactive and ligandable cysteines, the development of strategies that enable rapid cross-dataset comparisons between these studies represents an important opportunity for the cysteine chemoproteomics community that will enable a more comprehensive understanding of the cysteinome. Here we present CysDB as such a tool that unites high coverage chemoproteomic measures of identification, ligandability, and hyperreactivity across multiple studies, together with integration with relevant resources to provide metrics of functionality and disease relevance. CysDB achieves identification of an impressive 62,888 unique cysteines and 11,621 proteins, which represents a 100% increase in total number of identified cysteine residues compared with individual prior studies, with added potential for further growth as new data sets become available.

For our first step toward constructing CysDB, we accumulated and curated a selected set of cysteine chemoproteomics studies, which were prioritized because of the high coverage of identified cysteines. During our stringent data curation, we observed study-dependent differences in conventions for designating a cysteine as hyperreactive and/or ligandable. To account for the potential uncertainty caused by a general absence of field-wide data analysis conventions, we retained all hyperreactive and/or liganded cysteines to accurately represent each study's reported findings. The development of statistically rigorous conventions for the field will aid in normalizing future cross-dataset comparison efforts. Recently, in our studies we have required comparable ratios with low SDs identified across multiple biological replicates together with inclusion of inactive control datasets to further simplify removal of potentially spurious elevated ratios. For studies that rely on MS1-based quantification, so-called singleton values should be treated with an additional level of stringency, as these can prove more prone to yielding spurious ratios. These ratios are derived from peptides with precursor ions that have only been identified with either a heavy or light isotopic modification. Therefore, we followed general conventions for filtering singletons, by setting a maximum ratio value of $\log_2$(ratio) equivalent to 20 requiring identification of additional lower ratio ions. Future studies, including our own, will benefit significantly from harnessing advances in data acquisition and analysis to improve reproducibility, including imputation and data-independent acquisition, as showcased by recent efforts by the Wang group.[84]

Illustrating the utility of CysDB, we find that by combining datasets generated across multiple cell lines and using different labeling reagents, we substantially increase aggregate coverage of the cysteinome. Alongside cysteine coverage, CysDB reveals that cell line selection can impact not only which cysteines are identified in proteomes derived from different cell lines (**Figure S5**), but also the hyperreactivity and ligandability of individual cysteines. We ascribe these differences in part to both cell state specific expression as well as the stochastic nature of data dependent acquisition (DDA), which is the acquisition method used to generate nearly all datasets analyzed.

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

A key feature of CysDB is the inclusion of functional and disease annotations from UniProtKB, CGC, and ClinVar. We expect that the centralization of the annotations should allow rapid prioritization of ligandable cysteines for future studies. Showcasing the utility of cysteine chemoproteomics to access tough-to-drug classes of proteins, we find a considerable enrichment in transcription factors containing ligandable cysteines (**Figure 5C**). We also observe that many Cancer Gene Census driver genes contain a cysteine identified in a chemoproteomics study. These findings together with our observation that a smaller but still substantial 38% of all Cancer Gene Census genes contain a ligandable cysteine suggests opportunities for future studies to more comprehensively assess the ligandability of these genes.

During our efforts to map annotations generated from genomics data (e.g., ClinVar/Cancer Gene Census data), we encountered issues with mismapping for a subset of identifiers. While processing all datasets included in CysDB, we observed that a handful (16) of gene names did not map to UniProtKB protein accession numbers in a one-to-one manner during SQL querying; multiple HGNC or Gene Entrez symbols can be associated with a single protein identifier if the translated gene products are identical protein sequences.[26] Given the utility of a gene-centric search, we have incorporated such identifiers in this release of CysDB to aid future proteogenomic analysis.

An ongoing goal of CysDB is to facilitate expanding the scope of the ligandable and potentially druggable cysteinome, particularly for functional and disease-relevant proteins. Given our observed bias in CysDB ligandability datasets toward chloroacetamide and acrylamide moieties, we expect that future expansions of the ligandable cysteinome may stem in part from chemo proteomic studies using additional classes of electrophiles. In a similar manner, we expect that inclusion of datasets generated using alternatives to iodoacetamide as promiscuous cysteine reactive capping agents, including, for example, hypervalent iodine-based probes,[19] should further increase coverage of labeled cysteines. In this first iteration of CysDB, we have opted to restrict our datasets to those generated through lysate-based proteomic studies, which eliminates challenges associated with deconvolving changes in protein abundance from direct cysteine labeling. Given the importance of cell-based studies for target discovery and hit-to-lead optimization, we look forward to including such datasets in future releases, particularly when combined with bulk measures of protein abundance. In a similar manner, we look forward to incorporating redox proteomics datasets in subsequent iterations of CysDB, alongside generalized strategies to merge the diverse data formats generated by these studies. Looking ahead, we are enthusiastic about the continued growth of CysDB and encourage all interested users to consider submission of relevant chemoproteomics datasets that comply with our submission format (**Data S1**) and that include spectral files deposited in a public data repository, such as Pride.[85]

Acknowledgements

This study was supported by a Beckman Young Investigator Award (K.M.B.), DOD-Advanced Research Projects Agency (DARPA) grant D19AP00041 (K.M.B.), and NIGMS System and Integrative Biology grant 5T32GM 008185-33 (L.M.B.). We thank all members of the Backus lab for helpful suggestions. We thank S. Forli and J. Eberhardt for helpful suggestions.

Author Contributions

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

Declaration of Interests

K.M.B. is a paid consultant for Oncovalent Therapeutics and Matchpoint Therapeutics.

2.4 - Methods

Data and Code Availability

- This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.
- Original code has been deposited at https://github.com/lmboat/cysdb_app and is publicly available as of the date of publication.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Proteomics Data Analysis

Chemoproteomics data was collected from publicly accessible supplementary tables of previous literature.[2,4–11] Columns were parsed for UniProtKB protein identifiers and locations of the corresponding modified cysteine amino acid numbers to create a new identifier for CysDB: UniProtKBID_CYS#. Any cysteine classified as 'ligandable' or 'hyperreactive' is listed in CysDB as ligandable or hyperreactive. Individual ligandability and reactivity ratios found from each publication are listed in **Data S1** and **Data S2**. In some cases, for the ligandability and reactivity datasets, publications listed ratios for peptides simultaneously modified at multiple cysteines such as UniProtKBID_CYS#1_CYS#2, where the ratios provided for UniProtKBID_CYS#1_CYS#2 differed from UniProtKBID_CYS#1. Thus, ratios for peptides modified at multiple cysteines were not included in further analyses.

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

SMILES strings listed in the Supplementary Tables for each publication were copied and pasted into a new document. To obtain a uniform SMILES format for all the compounds in CysDB, published SMILES strings were converted into molecules and converted back into SMILES strings using RDKit.[53]

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

Cancer Gene Census (CGC) website reports were downloaded Sept. 2022 and mapped to CysDB data using UniProtKB accessions. Due to frequent UniProtKB updates, Gene symbols reported in the Cancer Gene Census were mapped to gene names in UniProtKB to identify the updated UniProtKB codes (2209-release).

Functional, Structural, and Druggability Annotations Data Analysis

Custom Python scripts classified protein functions based on annotations in the UniProtKB/Swiss-Prot[26] database (2209-release). UniProtKB accessions were collected from proteins with available ChEMBL and DrugBank UniProtKB annotations. Data from the Human Protein Atlas[29] (HPA) version 21.1 was downloaded and parsed to obtain genes targeted by FDA approved drugs. HGNC gene symbols were mapped to UniProtKB accessions to collect proteins targeted by FDA approved drugs.

Custom Python scripts classified protein functions based on annotations in the UniProtKB/Swiss-Prot database (2209-release), HPA version 21.1 and the ScaPD database.[86] UniProtKB keywords were parsed to classify proteins into five broad functional categories: chaperones/transporter/channel/receptor, enzyme, nucleic acid and small molecule binding, scaffolding/modulator/adaptor, transcription factor/regulator and uncategorized. Transcription factors, channels and transporters were also found using protein class descriptions from the HPA. In addition, examples of experimentally validated scaffolding proteins were collected from the ScaPD database. For proteins in more than one category, annotations were prioritized based on the following: enzyme > chaperones/transporter/channel/receptor > scaffolding/modulator/adaptor > transcription factor/regulator > nucleic acid and small molecule binding.

Counts of how many CysDB proteins had UniProtKB annotations for active sites, binding sites, catalytic activity, disulfide bonds and redox potentials were calculated based on matches between the position of the identified residue and UniProtKB functional annotation. Further parsing of UniProtKB active and binding site annotations were extracted to obtain specific residues and amino acid numbers. Positions of binding and active sites that were not cysteine residues were discarded. Exact amino acid positions of UniProtKB cysteine active and binding sites were cross-referenced with CysDB cysteine identifiers.

CysDB cysteines ‘in or near’ UniProtKB annotated active or binding sites were assessed using primary protein sequences. Positions of identified cysteines were found via their amino acid numbering. Annotated active or binding sites within +/-10 amino acids from the identified cysteine were considered as a cysteine ‘in or near’ an active or binding site.

Protein Data Bank[32] identifiers were found from UniProtKB annotations. Proteins without PDB structures were filtered out. PDB structures for proteins with PDB annotations were downloaded and parsed for amino acid numbering and residue names. A list of cysteines resolved in each PDB was stored for further processing. SIFTS[63] files, providing residue level mapping between PDB sequences and protein sequences, were downloaded for each PDB. Cysteines resolved in each PDB were mapped to their appropriate UniProtKB protein sequence and identifiers for PDB to UniProtKB pairs were created: PDB_C#_UniProtKBID_C#. From these paired identifiers, the number of unique UniProtKBID_C# records were counted to determine the number of UniProtKB cysteines resolved in PDBs.

CysDB cysteines ‘in or near’ UniProtKB annotated active or binding sites were assessed using 3D PDB structures. From the workflow described below (determining cysteines in PDB structures), PDB structures were parsed to find all neighboring residues within a 10 Angstrom distance of a cysteine residue. PDB_UniProtKB identifiers were created for each cysteine and corresponding list of neighboring residues. If the UniProtKB annotated active or binding sites were resolved in an associated crystal structure and found within the 10 Angstroms net, it was classified as a cysteine proximal to a known active or binding site.

CysDB Database

CysDB was created as a relational database using MySQL v.8.0. Overall, the database contains six tables and is hosted on Google Cloud. The major parent tables, ‘Datasets’ and ‘Identifiers’, were further broken down into child tables, such as ‘Ligandable’, ‘Reactive’, ‘Compound’ and ‘Warheads’ (**Figure S6**). The Datasets table contains information specific to each of the nine publications, while the Identifiers table contains information specific to each modified cysteine or protein identifier. Columns within Datasets and Identifiers include binary results for the following three categories: identified, hyperreactive and ligandable. However, individual competition ratios are listed in the Ligandable table and individual reactivity ratios are listed in the Reactive table. Calculated molecular properties for ‘drug-likeness’ were acquired using RDKit[53] and are stored in the ‘Compounds’ table. This table also contains the CysDB compound identifier mapped to their associated publication abbreviation or designated name. Group compound identifiers (‘‘GROUP_WARHEAD_#’’) were defined by unique standardized SMILES strings and individual compound identifiers (‘‘WARHEAD_#’’) were defined by unique standardized SMILES string, cell line and publication author combinations. Finally, the warhead table holds chemotype classifications for each compound. The five chemotype classifications were as follows: acrylamide, bromoacetamide, chloroacetamide, dimethyl fumarate and other.

CysDB Web Application

The CysDB web application was developed using the Shiny R package (https://shiny.rstudio.com/). Schematics of protein sequence chains, domains and motifs on the CysDB web server are constructed using the drawProteins R package (https://github.com/ brennanpincardiff/drawProteins). Interactive viewing of PDB crystal structures is performed using NGLViewR (https://github.com/ nglviewer/nglview). Protein protein interaction networks are accessed via the STRING database (https://string-db.org/). Gene set library enrichment analyses are provided with the Enrichr R package (https://maayanlab.cloud/Enrichr/) and ontology enrichment plots are produced with the gprofiler2 R package (https://biit.cs.ut.ee/gprofiler/gost). All plots are generated with the ggplot2 and plotly (https://plotly.com/r/) R libraries.

Quantification and Statistical Analysis

Enrichment of Panther 2016, Pfam Domains 2019 and OMIM Disease gene set library terms were performed using the GSEApy package.[87] Proteins identified by chemoproteomics studies in CysDB were utilized as the background protein set. UniProtKB protein identifiers were mapped to Entrez gene symbols as input for Enrichr. P-values were computed from Fisher's exact test to determine the significance of each enriched term. The negative log of these p-values was calculated using R.

Additional Resources

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

Dataset Addition to CysDB Guidelines

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

2.5 - Supporting Information

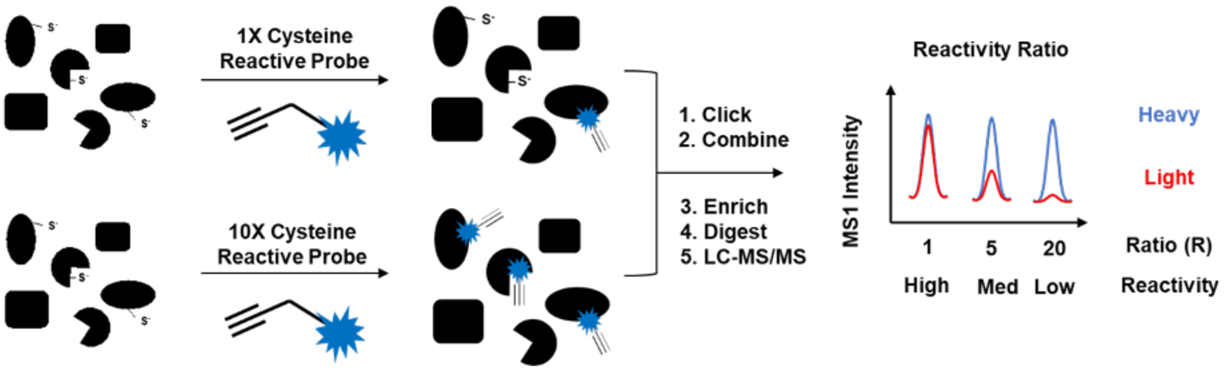


Figure S1. General chemoproteomics workflow for measuring intrinsic cysteine-reactivity towards electrophilic probes such as iodoacetamide alkyne (IAA), related to **Figure 1**. For these intrinsic reactivity studies, including those that use the Isotopic Tandem Orthogonal Proteolysis–ABPP (isoTOP–ABPP) platform, relative cysteine labeling by a high (10x) and low (1x) concentration of IAA or other probes are compared using isotopically labeled enrichment handles and MS1-based quantification. Hyper-reactive residues are those that show R10:1 ratios close to 1, indicating saturation of labeling at the lower reagent concentration.

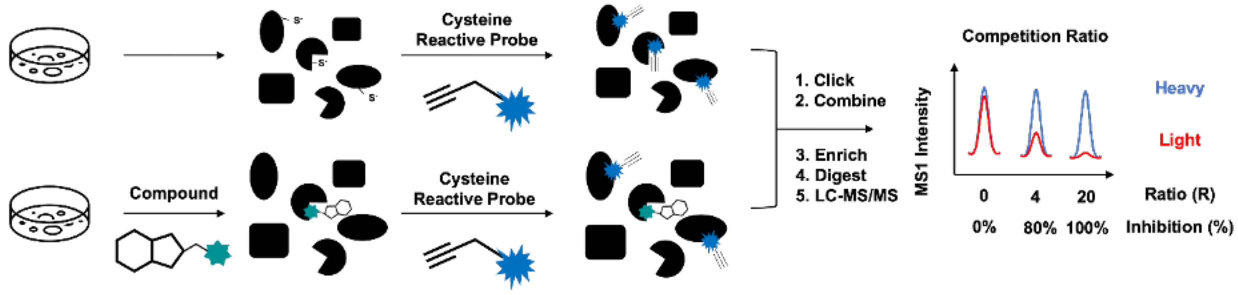


Figure S2. General chemoproteomics workflow for measuring cysteine ligandability using competitive isoTOP-ABPP and related methods, related to **Figure 1**. Proteomes are treated with electrophilic compounds or vehicle (DMSO), labeled with an iodoacetamide (IA)-alkyne probe, and conjugated to isotopically-differentiated, biotin enrichment handles by click chemistry. Treated and control samples are combined, processed, and analyzed by LC-MS/MS, where the isotopic label is used to distinguish between peptides from control and fragment-treated samples, with elevated RH:L ratios indicative of a liganded cysteine.

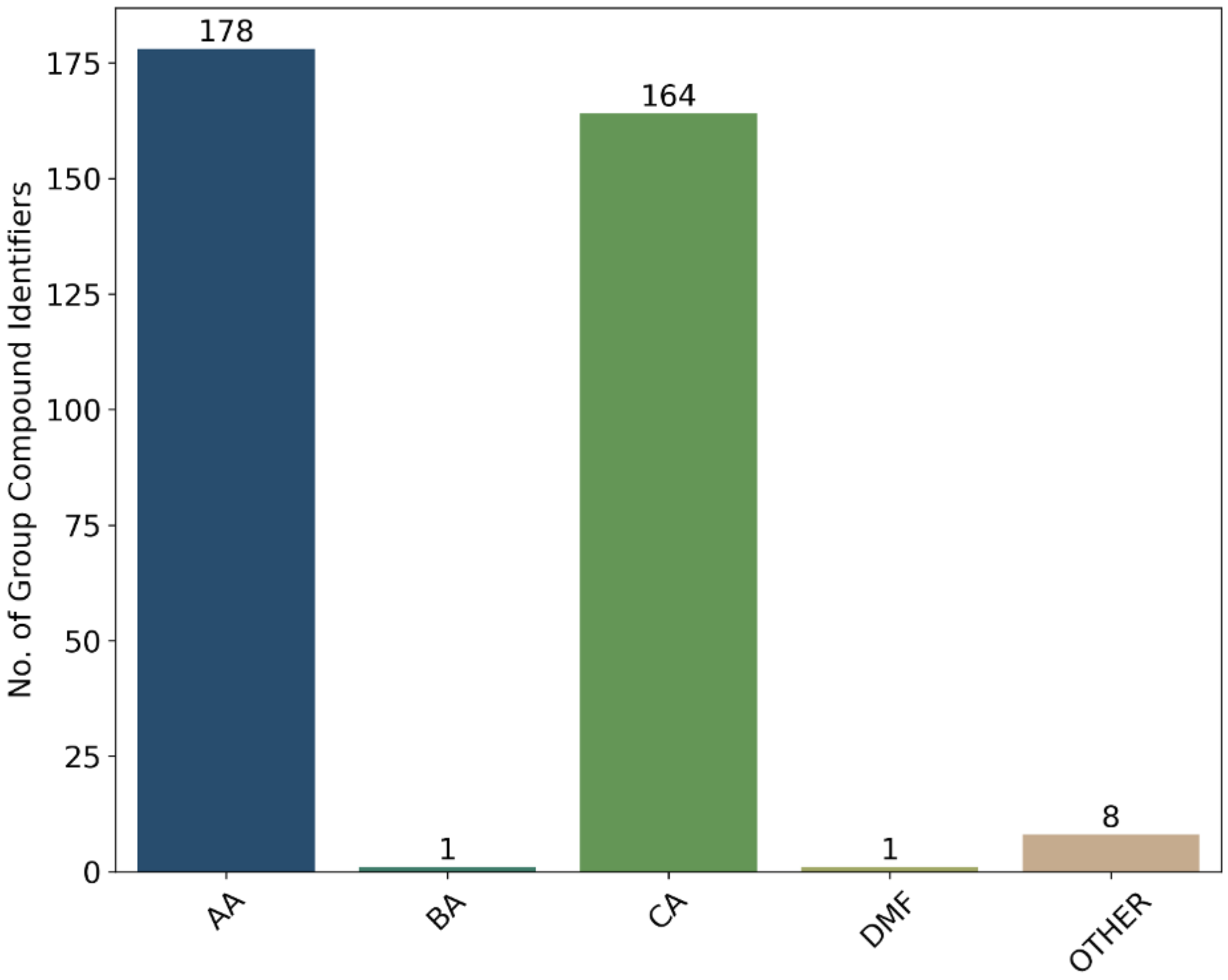


Figure S3. Total number of unique compounds for each warhead category in CysDB Lig, related to **Figure 1**: acrylamide (AA), bromoacetamide (BA), chloroacetamide (CA), dimethyl fumarate (DMF) and other (OTHER). Unique compounds were determined by SMILES strings/group compound identifiers.

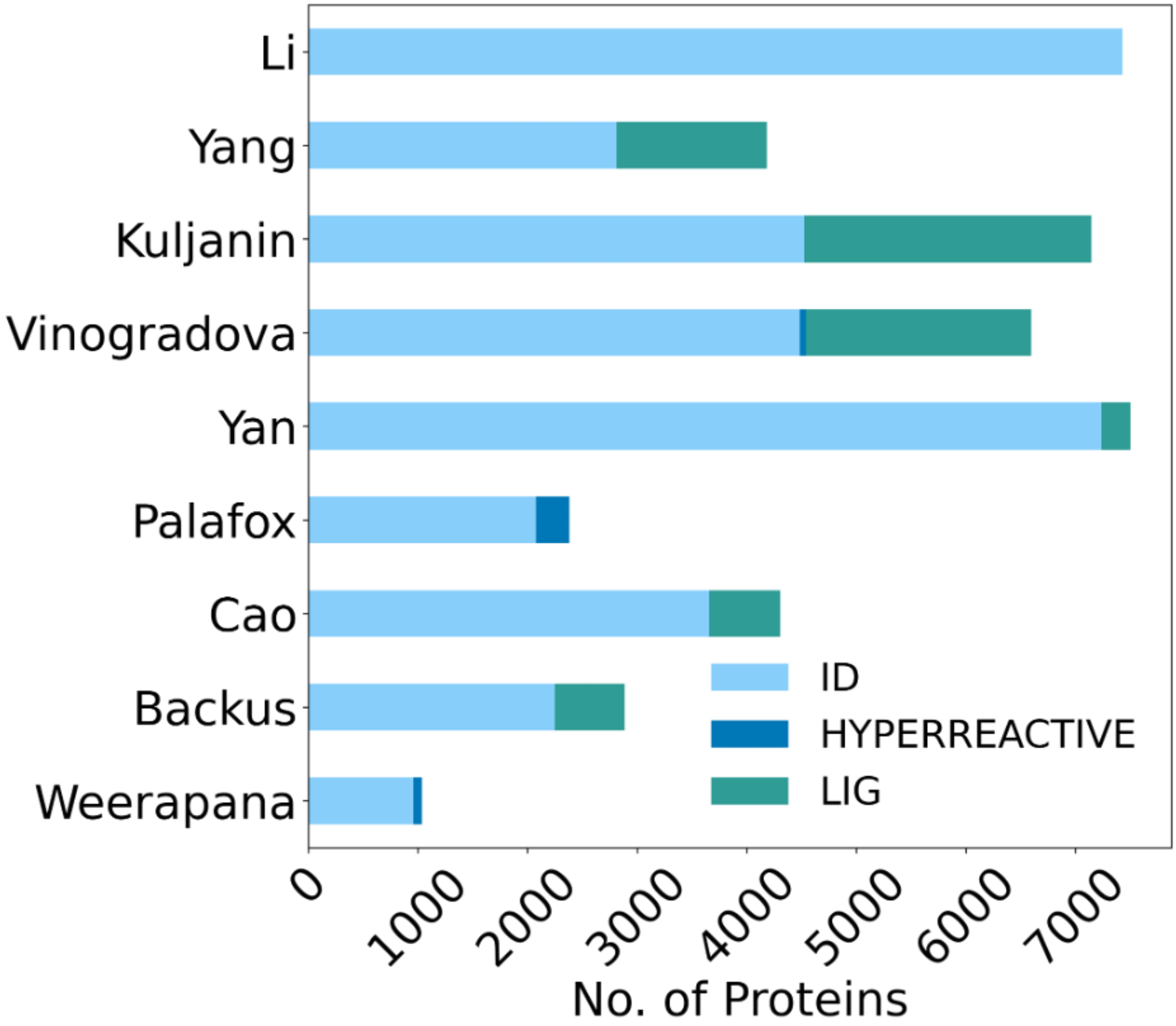
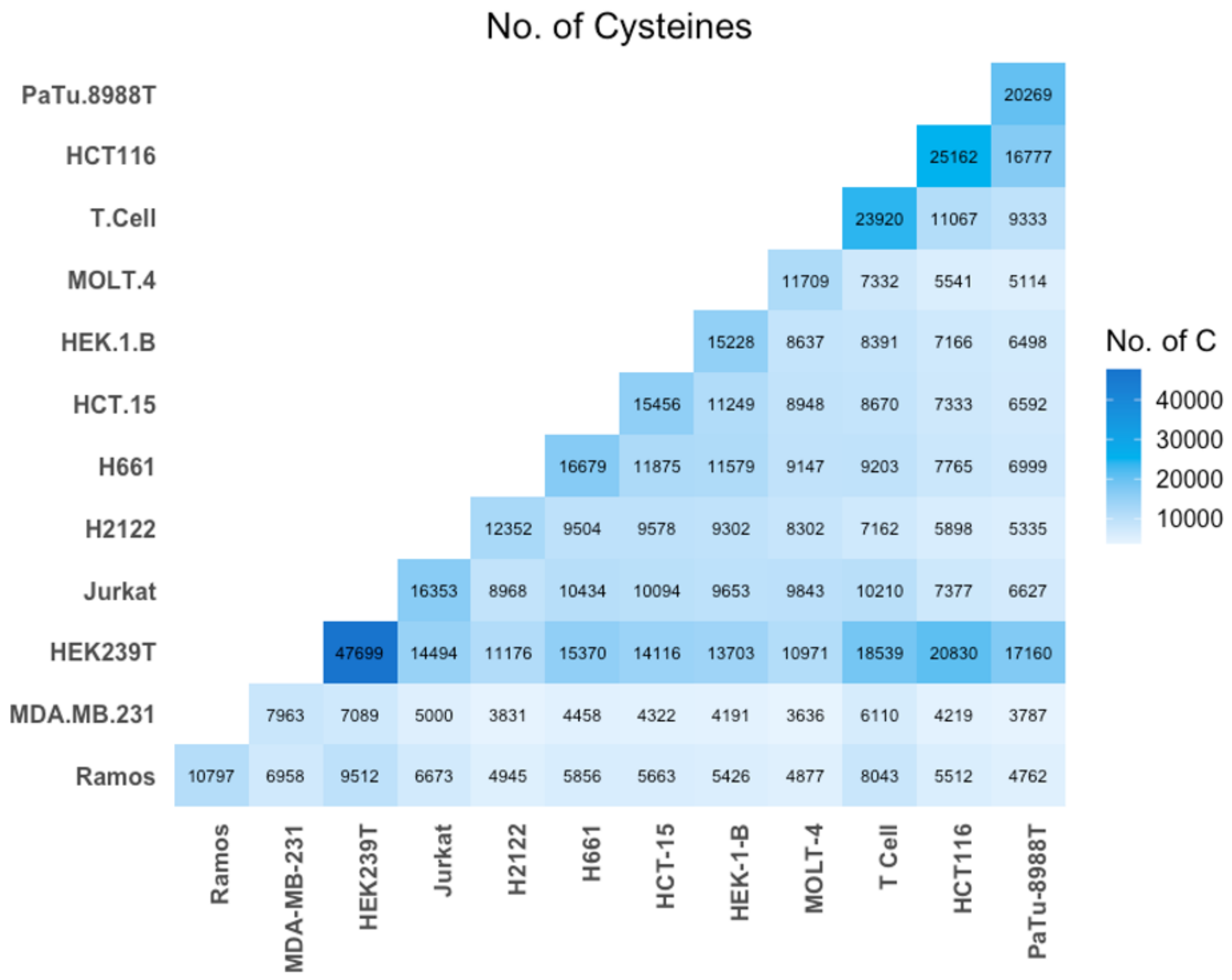


Figure S4. Total number proteins for each category in CysDB & in UniProtKB/Swiss-Prot, related to **Figure 1**.



	Ramos	MDA-MB-231	HEK239T	Jurkat	H2122	H661	HCT-15	HEK-1-B	MOLT-4	T Cell	HCT116	PaTu-8988T
PaTu.8988T												20269
HCT116											25162	16777
T.Cell										23920	11067	9333
MOLT.4									11709	7332	5541	5114
HEK.1.B								15228	8637	8391	7166	6498
HCT.15							15456	11249	8948	8670	7333	6592
H661						16679	11875	11579	9147	9203	7765	6999
H2122					12352	9504	9578	9302	8302	7162	5898	5335
Jurkat				16353	8968	10434	10094	9653	9843	10210	7377	6627
HEK239T			47699	14494	11176	15370	14116	13703	10971	18539	20830	17160
MDA.MB.231		7963	7089	5000	3831	4458	4322	4191	3636	6110	4219	3787
Ramos	10797	6958	9512	6673	4945	5856	5663	5426	4877	8043	5512	4762

Figure S5. Number of identified cysteines shared between different cell lines, related to **Figure 1**.

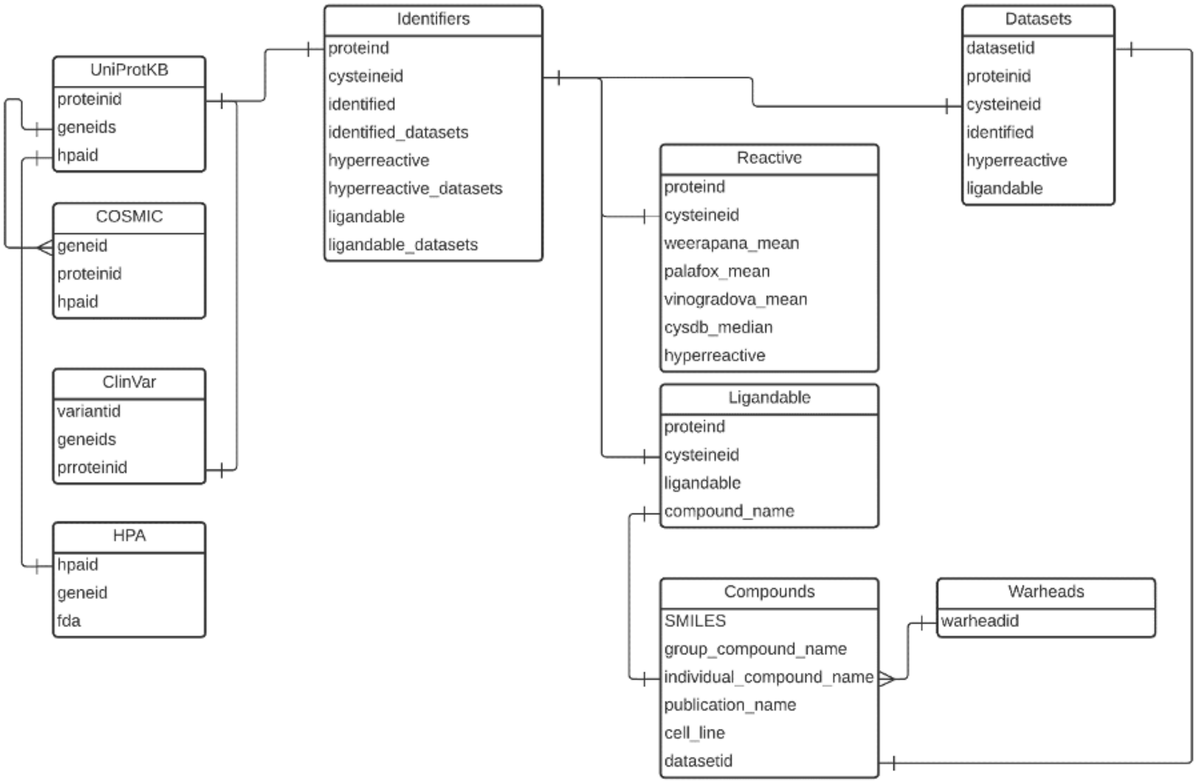


Figure S6. Entity-relationship diagram of all ten tables in CysDB and relationships with external data sources, such as UniProtKB, COSMIC, ClinVar and the Human Protein Atlas (HPA), related to **Figure 2**.

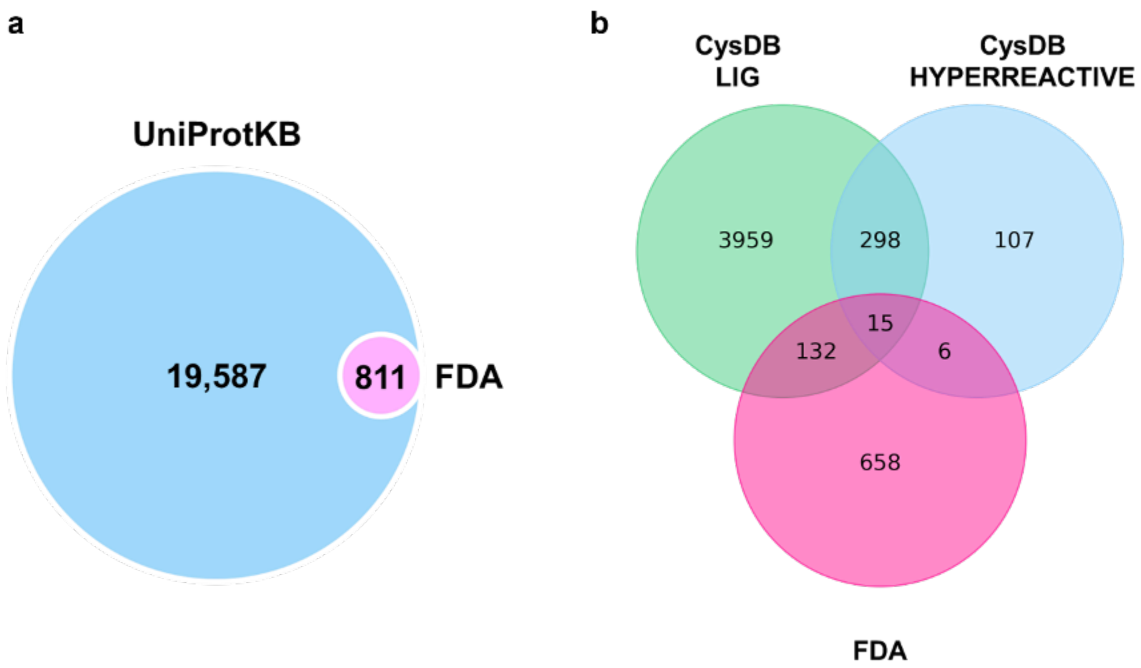


Figure S7. Total number of proteins in the human proteome from UniProtKB/Swiss-Prot and the subset targeted by FDA approved drugs (a). Total number of CysDB ligandable proteins, CysDB hyperreactive proteins and proteins targeted by FDA approved drugs (b), related to **Figure 4**.

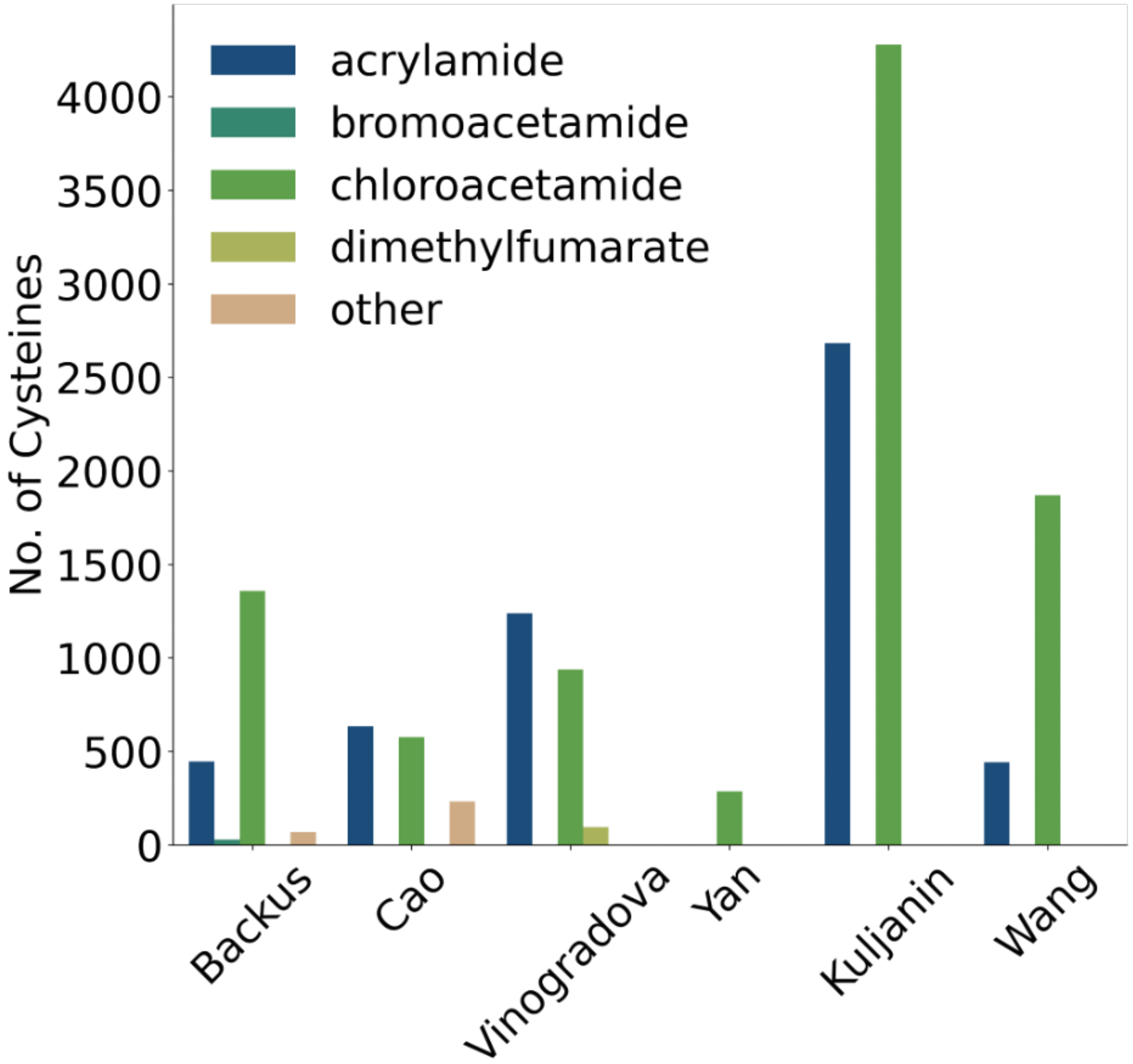


Figure S8. Total number of cysteines with an R > 4 by each warhead per dataset, in aggregate across all cell lines analyzed, related to **Figure 4**.

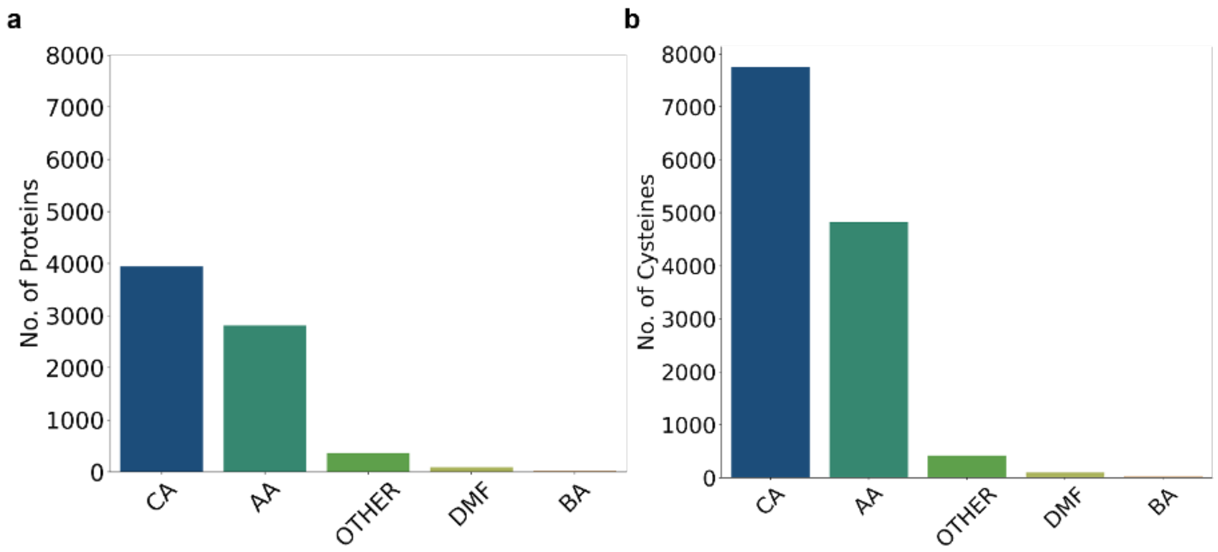


Figure S9. Total number of proteins (a) and cysteines (b) liganded by the following electrophiles, related to **Figure 4**: chloroacetamides (CA), acrylamides (AA), other (OTHER), dimethyl fumarate (DMF) and bromoacetamides (BA). Note, some proteins or cysteines were liganded by multiple warheads. Therefore, the counts in these graphs are not reflective of mutually exclusive events.

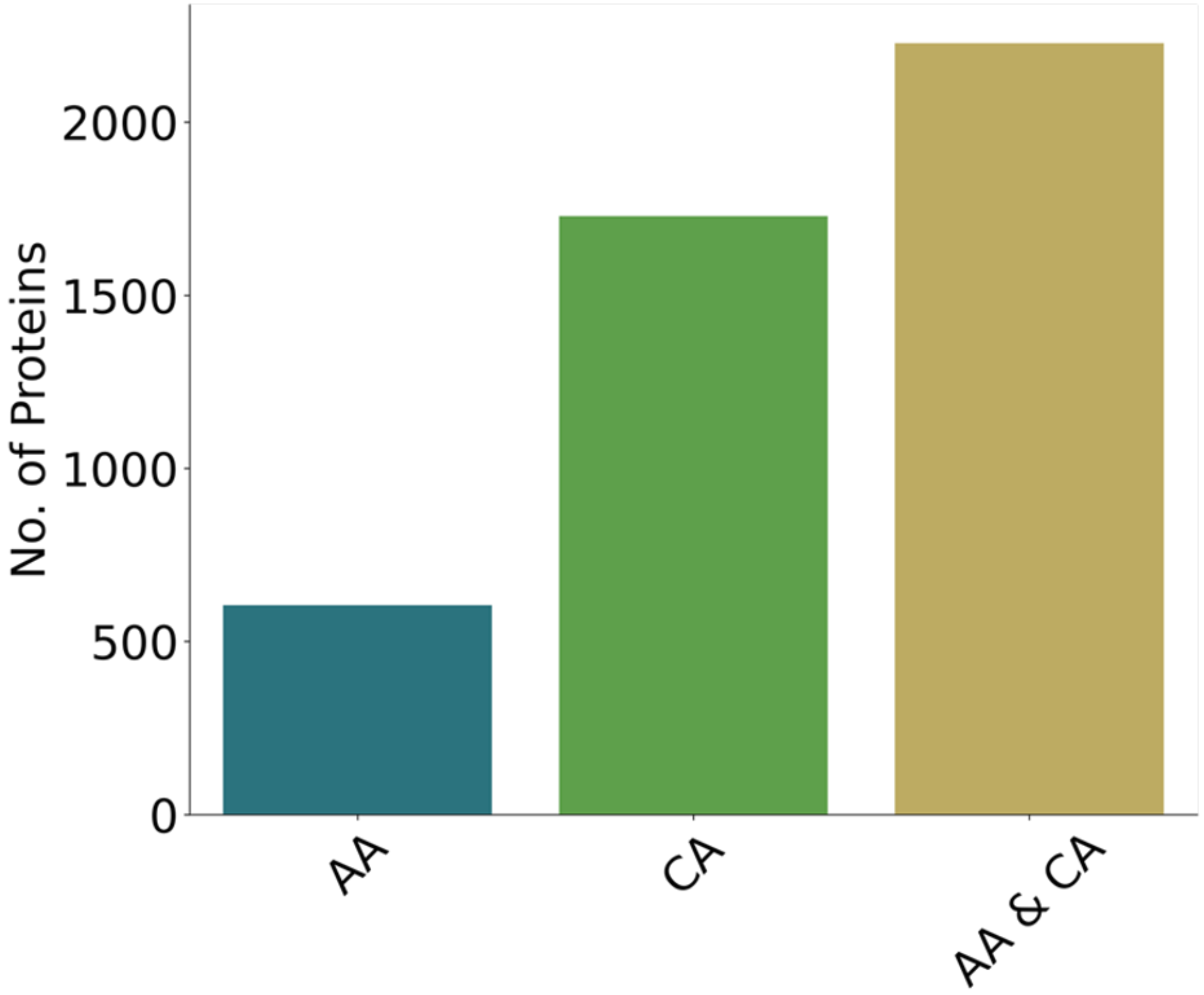


Figure S10. Total number of proteins liganded by both acrylamides and chloroacetamides, exclusively acrylamides and exclusively chloroacetamides, related to **Figure 4**.

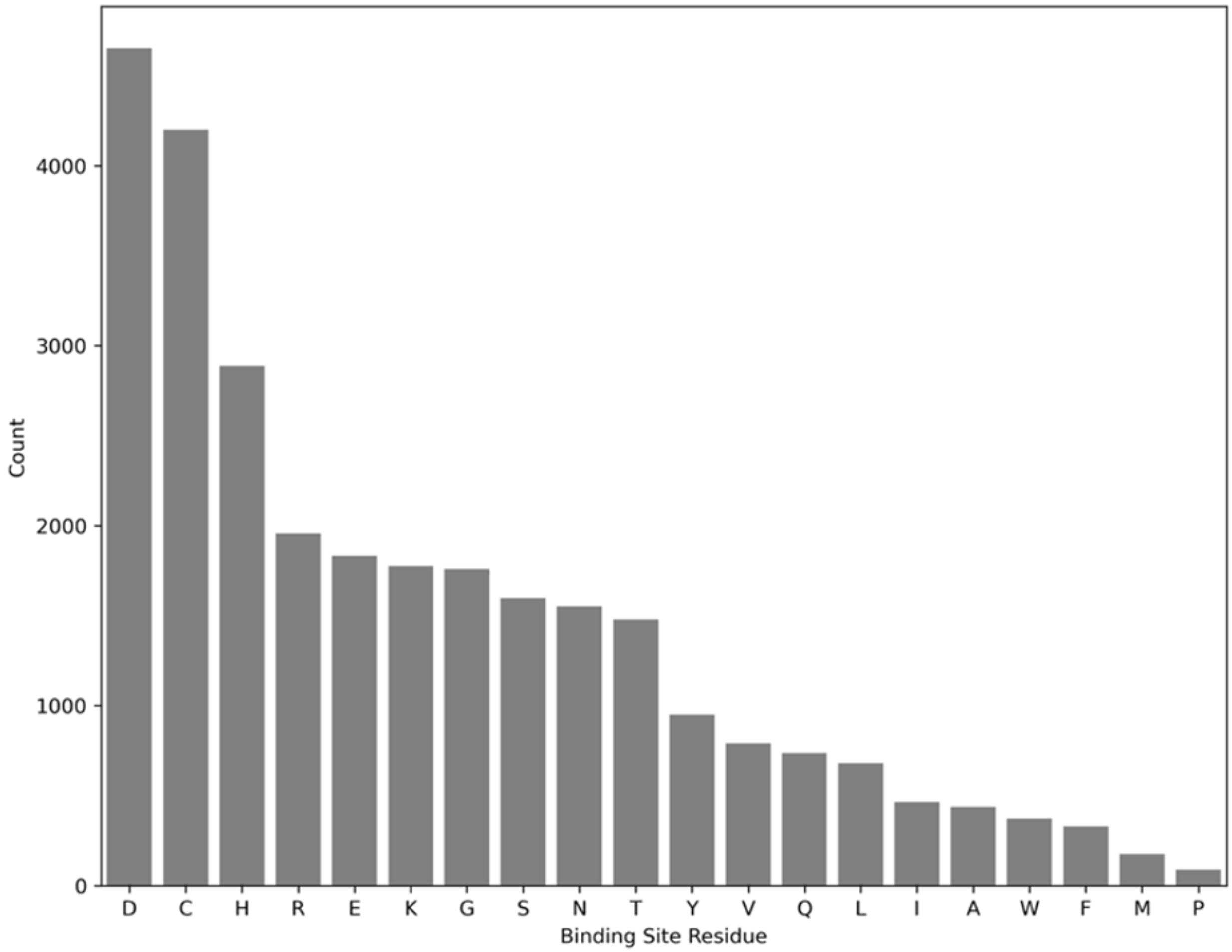


Figure S11. Distribution of amino acids annotated as binding sites in UniProtKB proteins, related to **Figure 5**.

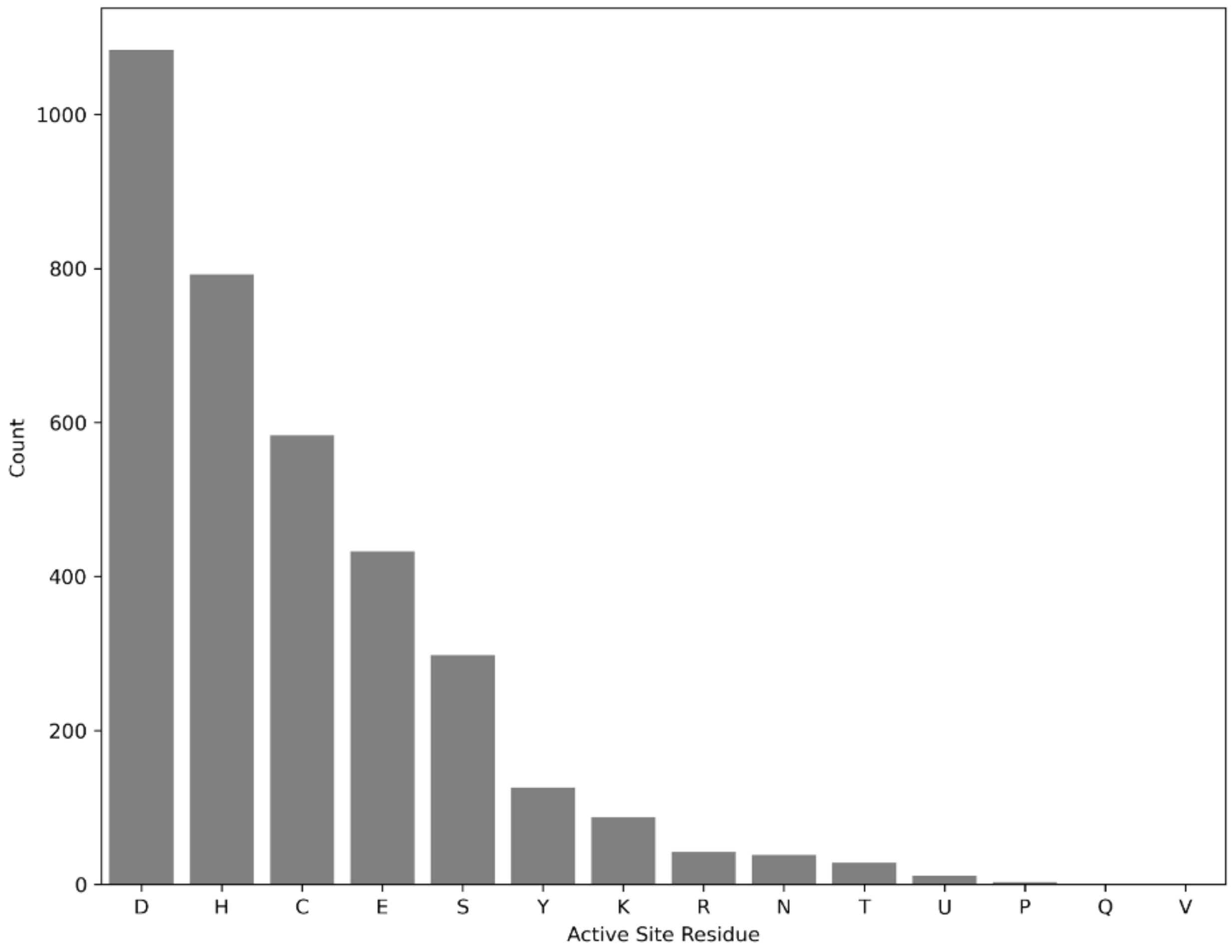


Figure S12. Distribution of amino acids annotated as active sites in UniProtKB proteins, related to **Figure 5**.

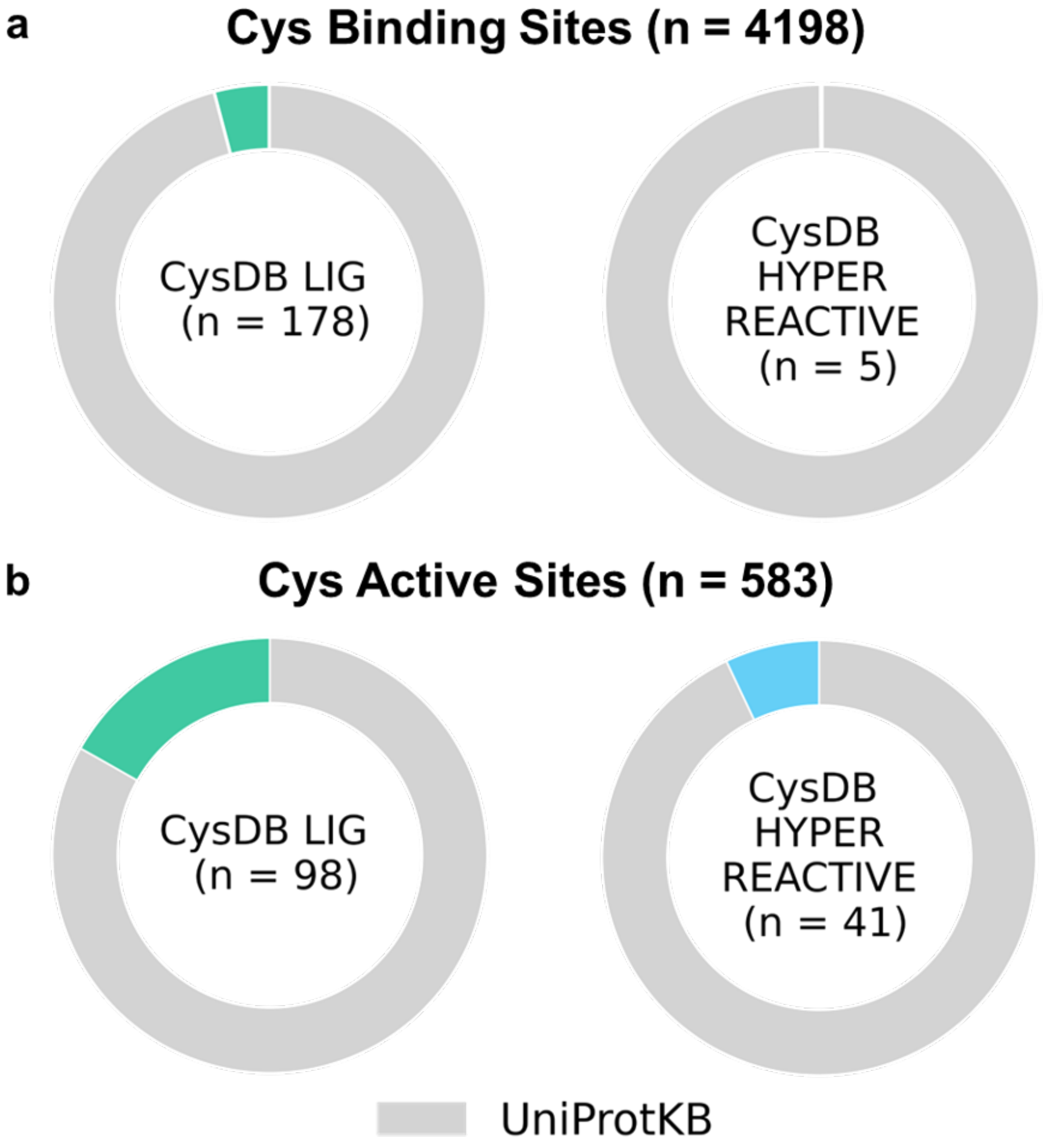


Figure S13. Distributions of ligandable (green) and hyperreactive (light blue) cysteines annotated as cysteine-specific binding sites (a) or cysteine-specific active sites (b), related to **Figure 5**. The total number of cysteines in UniProtKB annotated as binding or active sites are shown in gray.

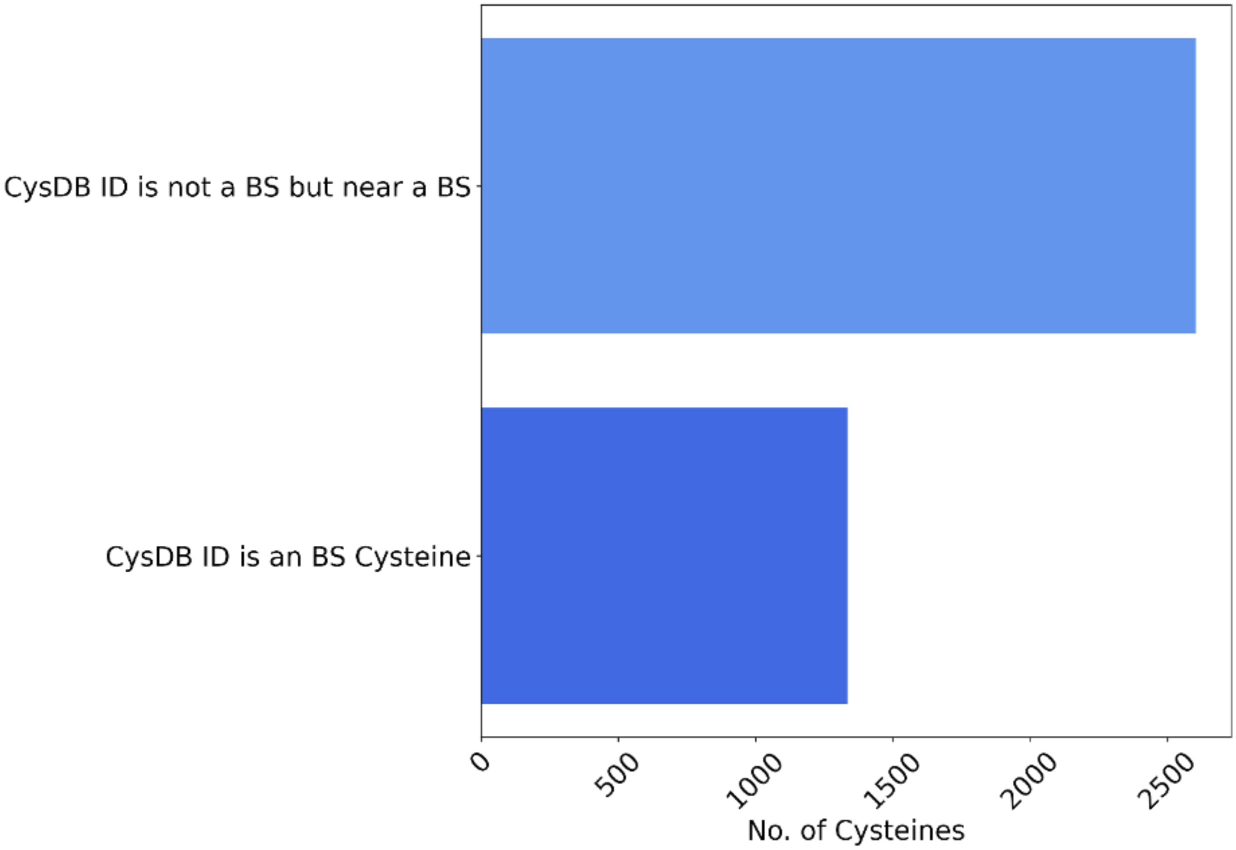


Figure S14. Number of CysDB ID cysteines that are annotated binding sites (BS) and cysteines that are not binding sites but in or near a binding site in 1D sequence, related to **Figure 5**. Primary sequences were searched +/- 10 amino acids from the location of a detected cysteine. If another binding site was within this +/- 10 amino acid window, the cysteine was considered 'in or near' a binding site.

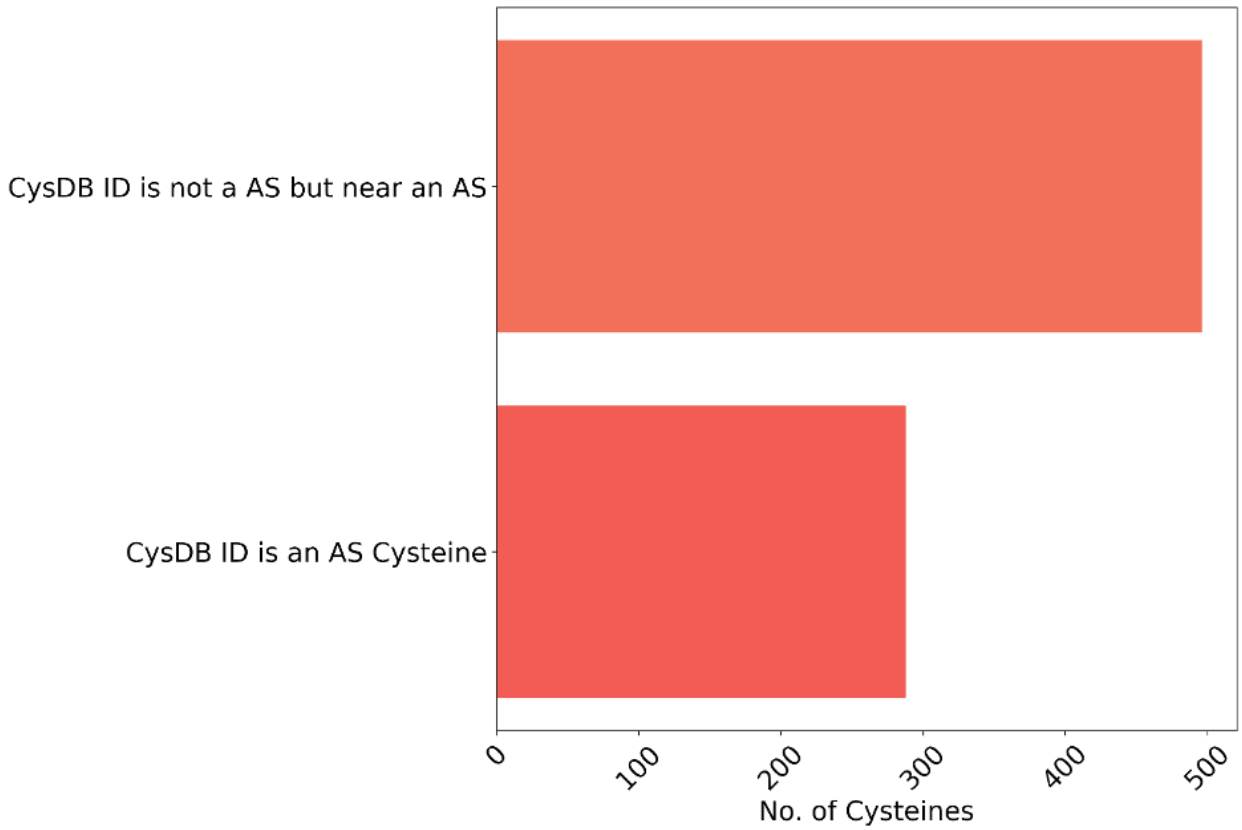


Figure S15. Number of CysDB ID cysteines that are annotated active sites (AS) and cysteines that are not active sites but in or near an active site in 1D sequence, related to **Figure 5**. Primary sequences were searched +/- 10 amino acids from the location of a detected cysteine. If another active site was within this +/- 10 amino acid window, the cysteine was considered 'in or near' an active site.

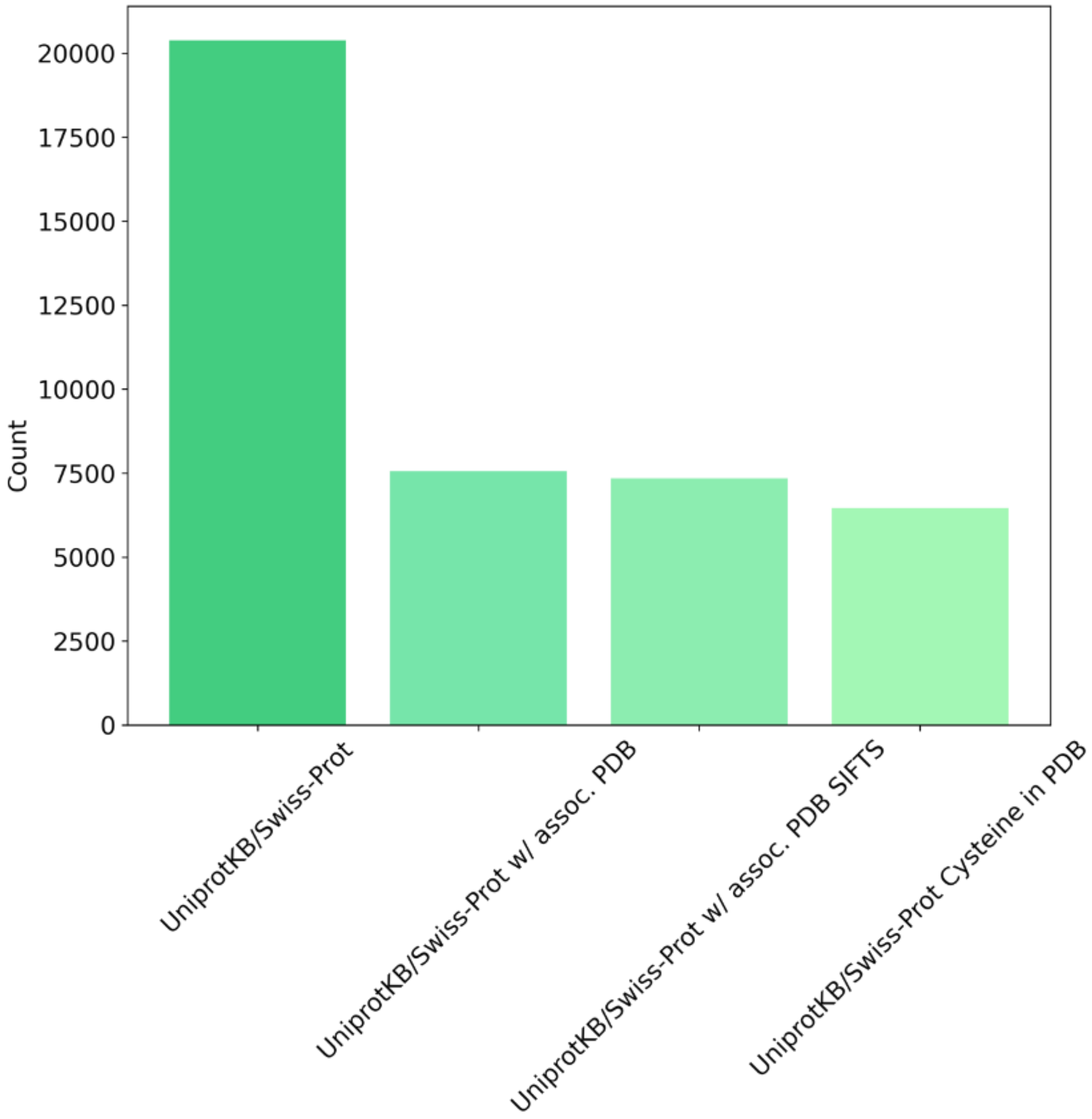


Figure S16. Number of UniProtKB proteins in the human proteome, with an associated PDB structure, residue mapped SIFTS file and with a cysteine resolved in the corresponding associated PDB, related to **Figure 5**.

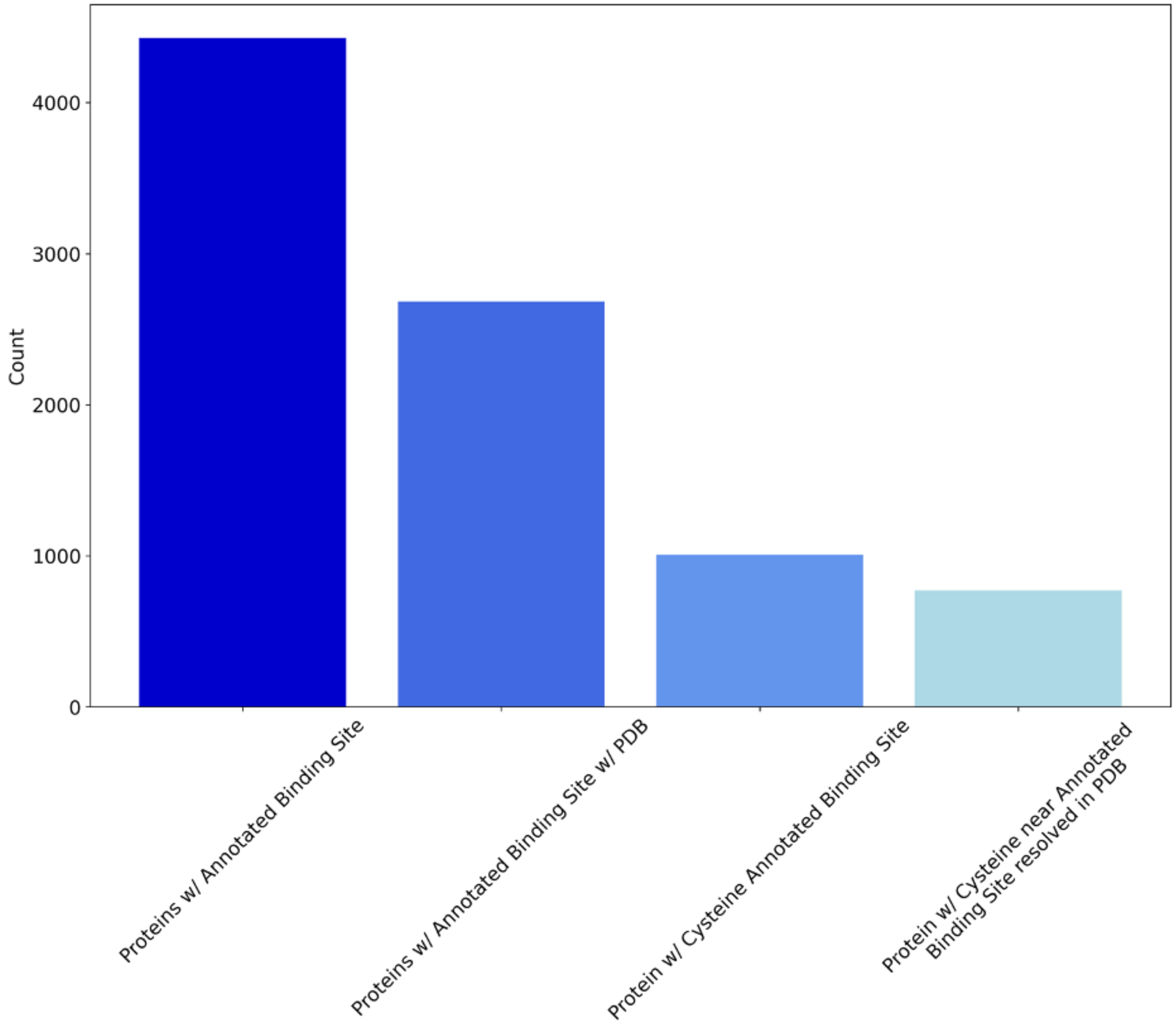


Figure S17. Number of UniProtKB proteins with an annotated binding site, associated PDB structure, with an annotated cysteine binding site and with cysteines near an annotated binding site in an associated PDB structure, related to **Figure 5** and see **STAR Methods**. The distance from the sulfur atom of each cysteine to an annotated binding site residue was calculated. Cysteines within 10 Angstroms of the annotated binding site residue were considered as cysteines 'in or near' binding sites.

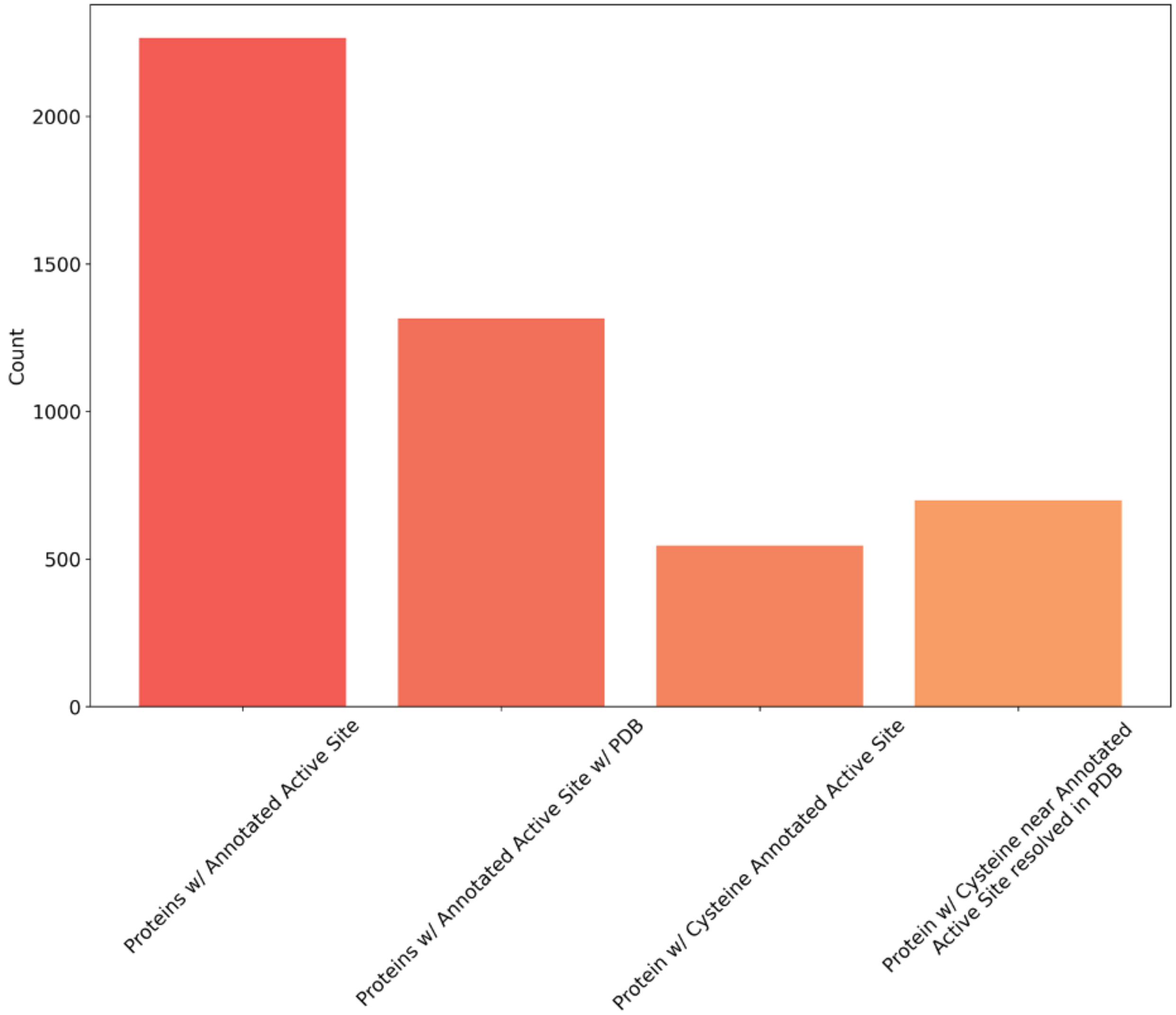


Figure S18. Number of UniProtKB human proteins with an annotated active site, associated PDB structure, with an annotated as cysteine active site and with cysteines near an annotated active site in an associated PDB structure, related to **Figure 5** and see **STAR Methods**. The distance from the sulfur atom of each cysteine to an annotated active site residue was calculated. Cysteines within 10 Angstroms of the annotated active site residue were considered as cysteines 'in or near' active sites.

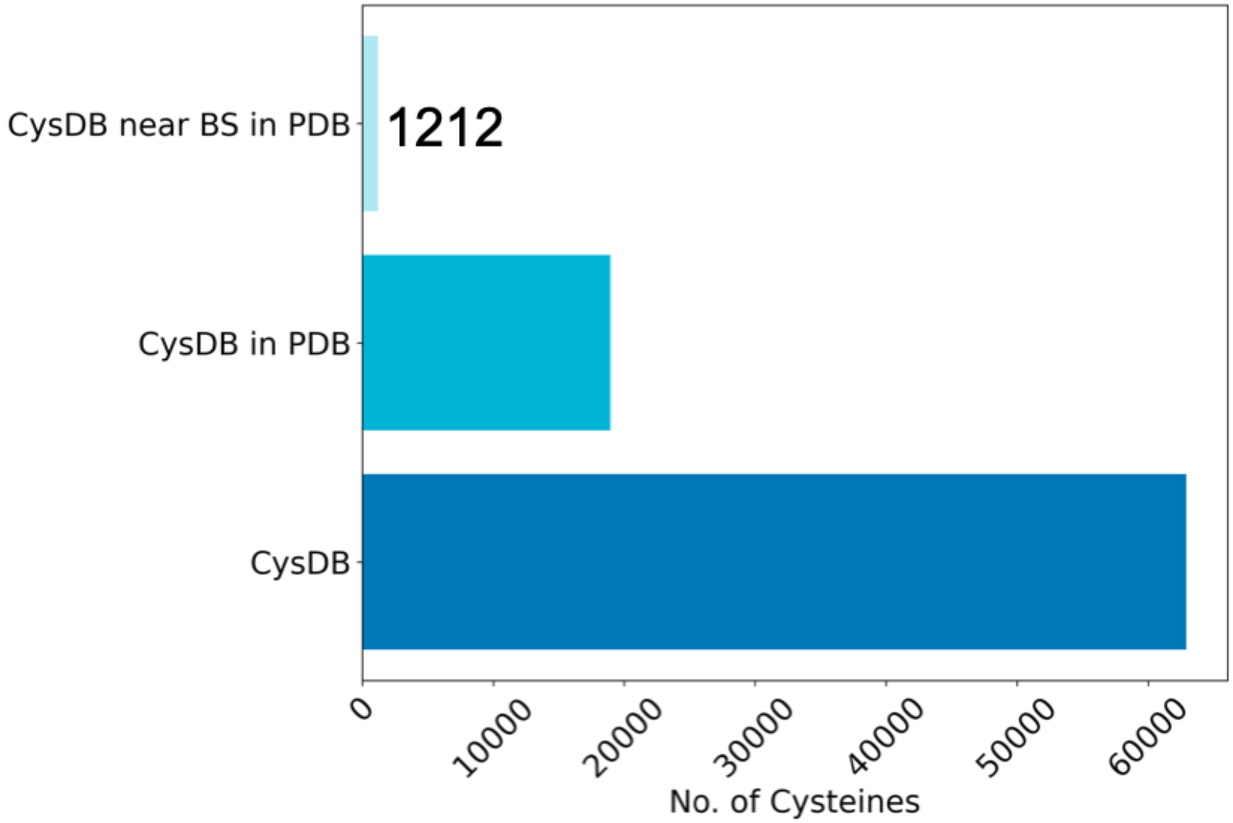


Figure S19. Number of CysDB ID cysteines identified by chemoproteomics, resolved in an associated PDB and CysDB ID cysteines that are not annotated binding sites but are 'in or near' an annotated binding site in 3D space, related to **Figure 5**. Proteins with an annotated binding site, annotated as a binding site resolved in an associated PDB structure and with cysteines 'in or near 'an annotated binding site. The distance from the sulfur atom of each cysteine to an annotated binding site residue was calculated. Cysteines within 10 Angstroms of the annotated binding site residue were considered as cysteines 'in or near' binding sites.

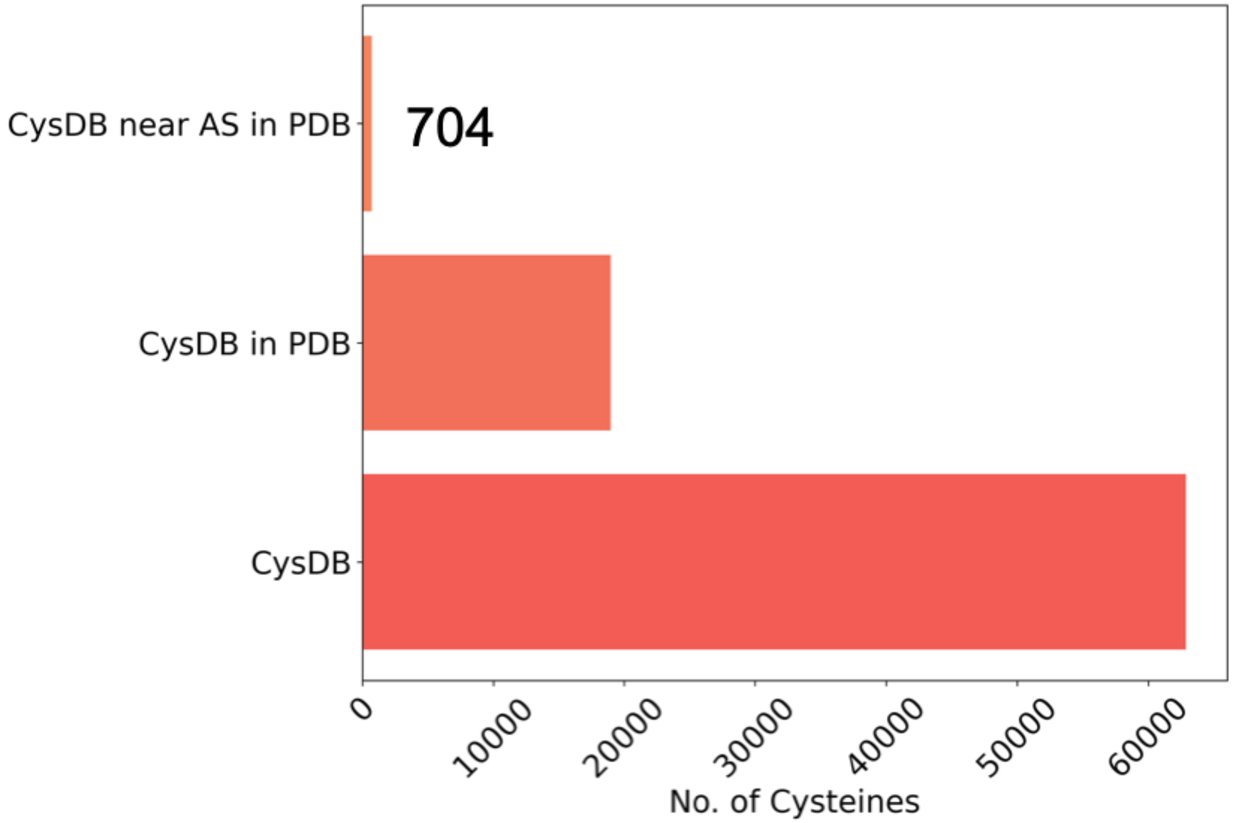


Figure S20. Number of CysDB ID cysteines identified, resolved in an associated PDB and CysDB ID cysteines that are not annotated active sites but are 'in or near' an annotated active site in 3D space, related to **Figure 5**. Proteins with an annotated active site, annotated as an active site resolved in an associated PDB structure and with cysteines 'in or near' an annotated binding site. The distance from the sulfur atom of each cysteine to an annotated active site residue was calculated. Cysteines within 10 Angstroms of the annotated active site residue were considered as cysteines 'in or near' active sites.

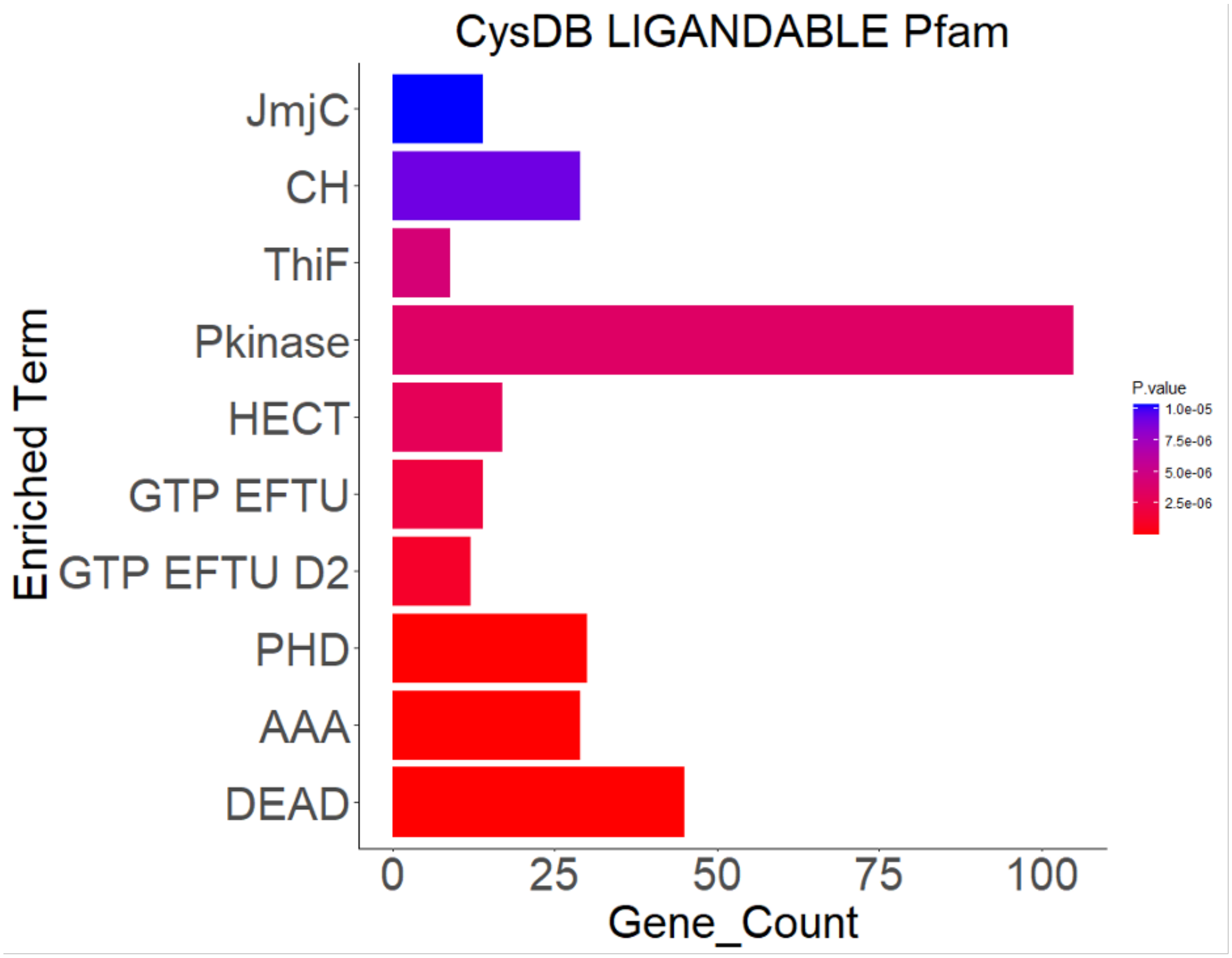


Figure S21. Top 10 enriched protein domains from Pfam-term enrichment analysis of liganded proteins with gene counts, related to **Figure 6**.

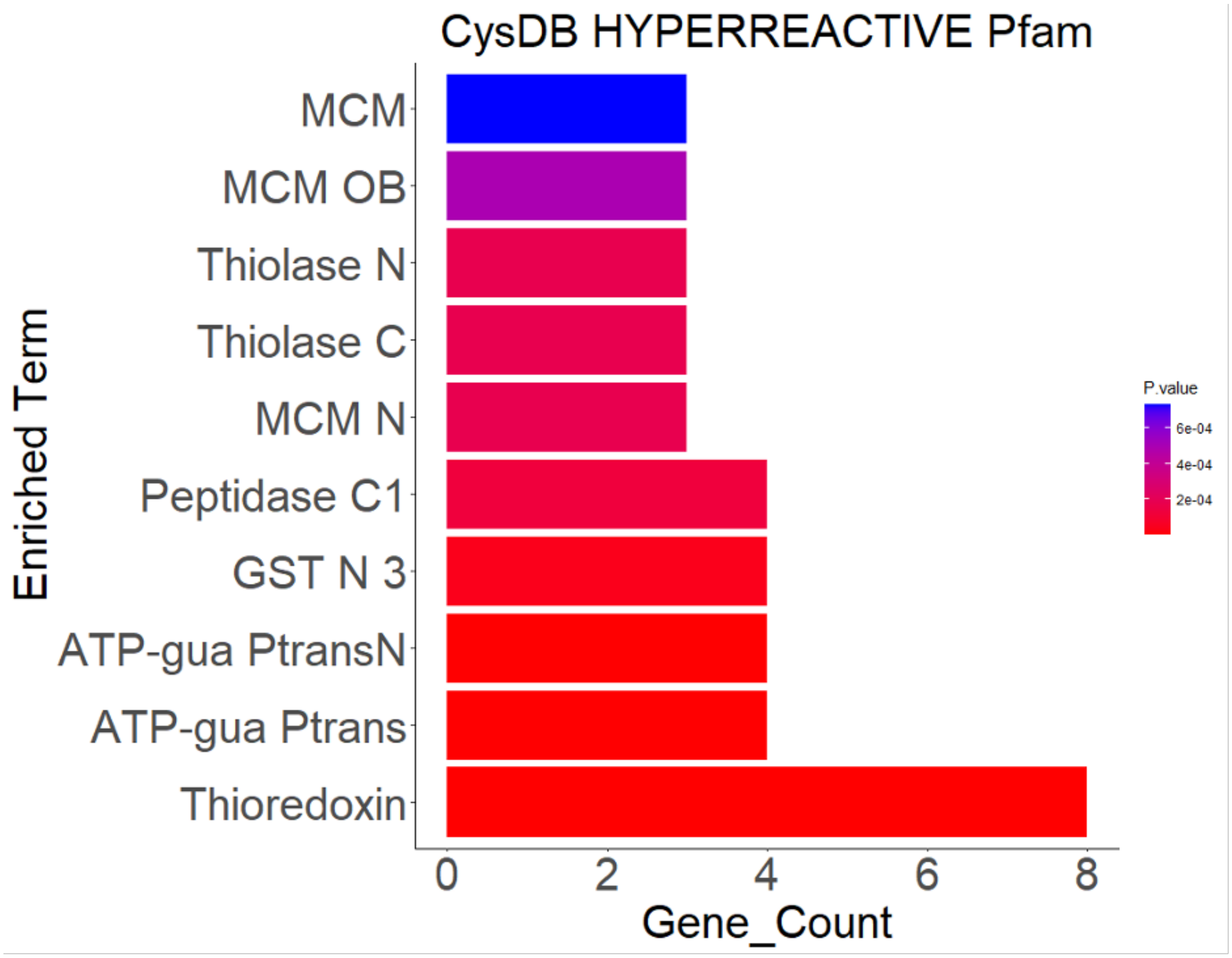


Figure S22. Top 10 enriched protein domains from Pfam-term enrichment analysis of hyper-reactive proteins with gene counts, related to **Figure 6**.

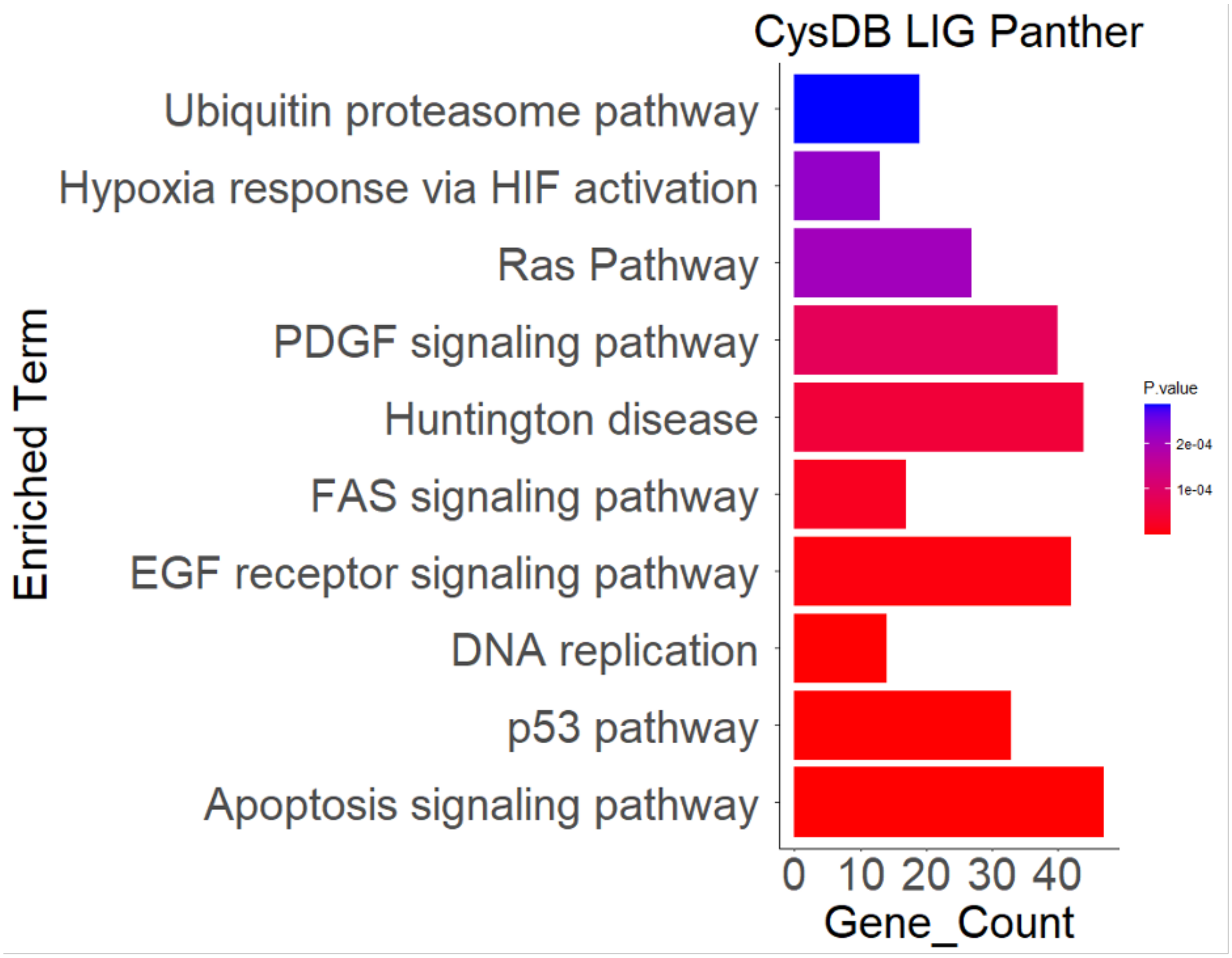


Figure S23. Top 10 enriched pathways from Panther-term enrichment analysis of liganded proteins with gene counts, related to **Figure 6**.

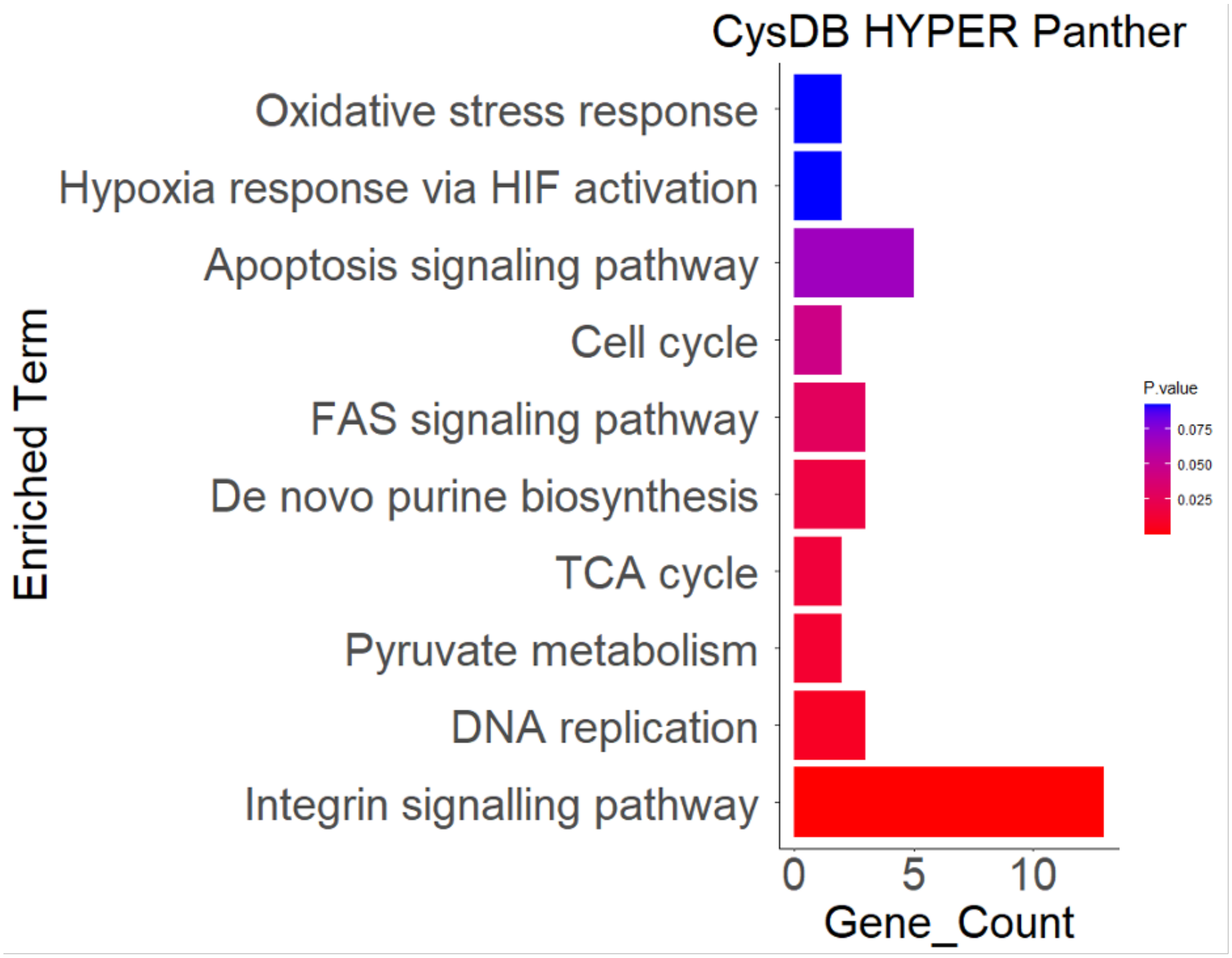


Figure S24. Top 10 enriched pathways from Panther-term enrichment analysis of hyperreactive proteins with gene counts, related to **Figure 6**.

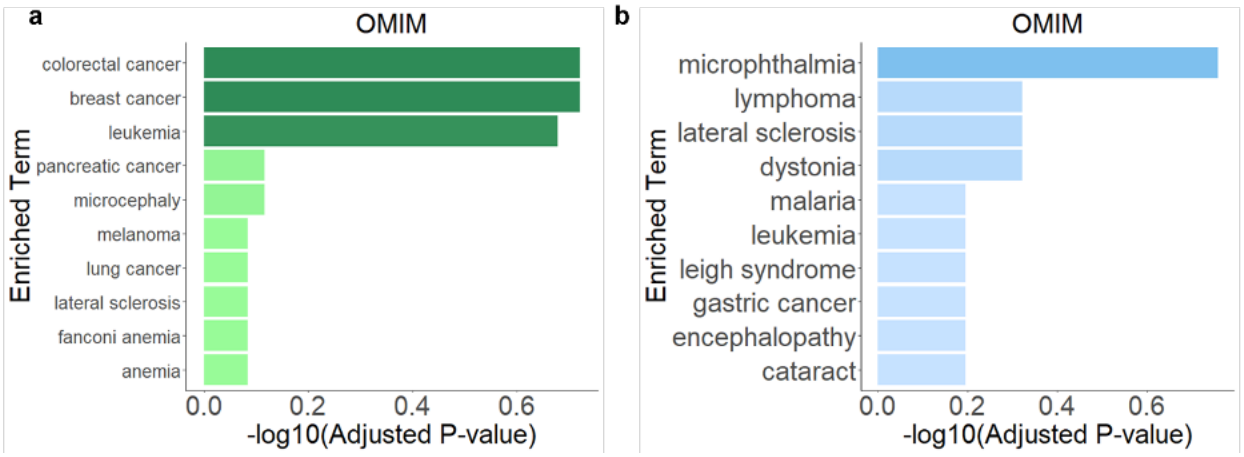


Figure S25. Top 10 enriched pathways from OMIM-term enrichment analysis of ligandable proteins (a) and hyperreactive proteins (b), related to **Figure 6**.

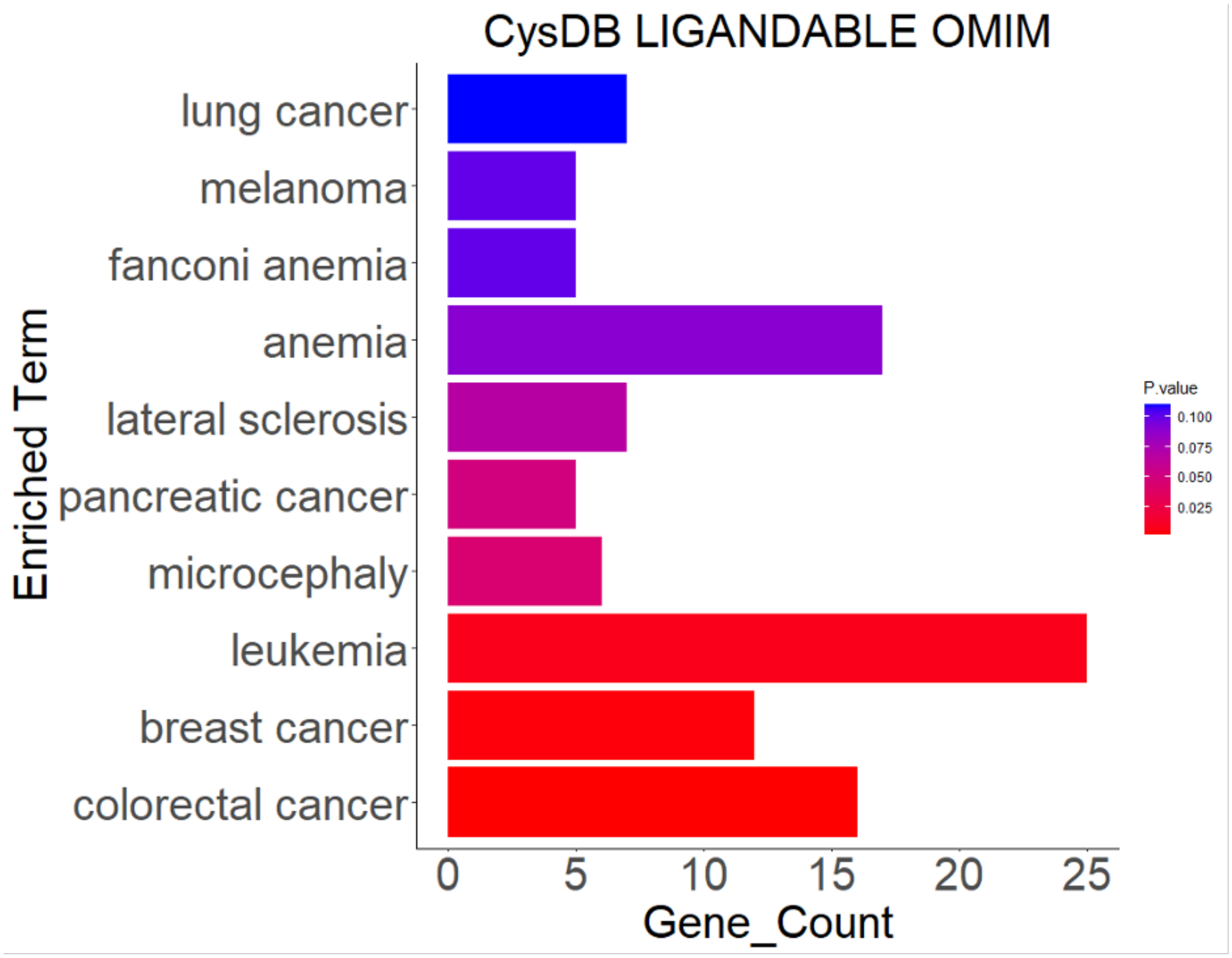


Figure S26. Top 10 enriched pathways from OMIM-term enrichment analysis of ligandable proteins with gene counts, related to **Figure 6**.

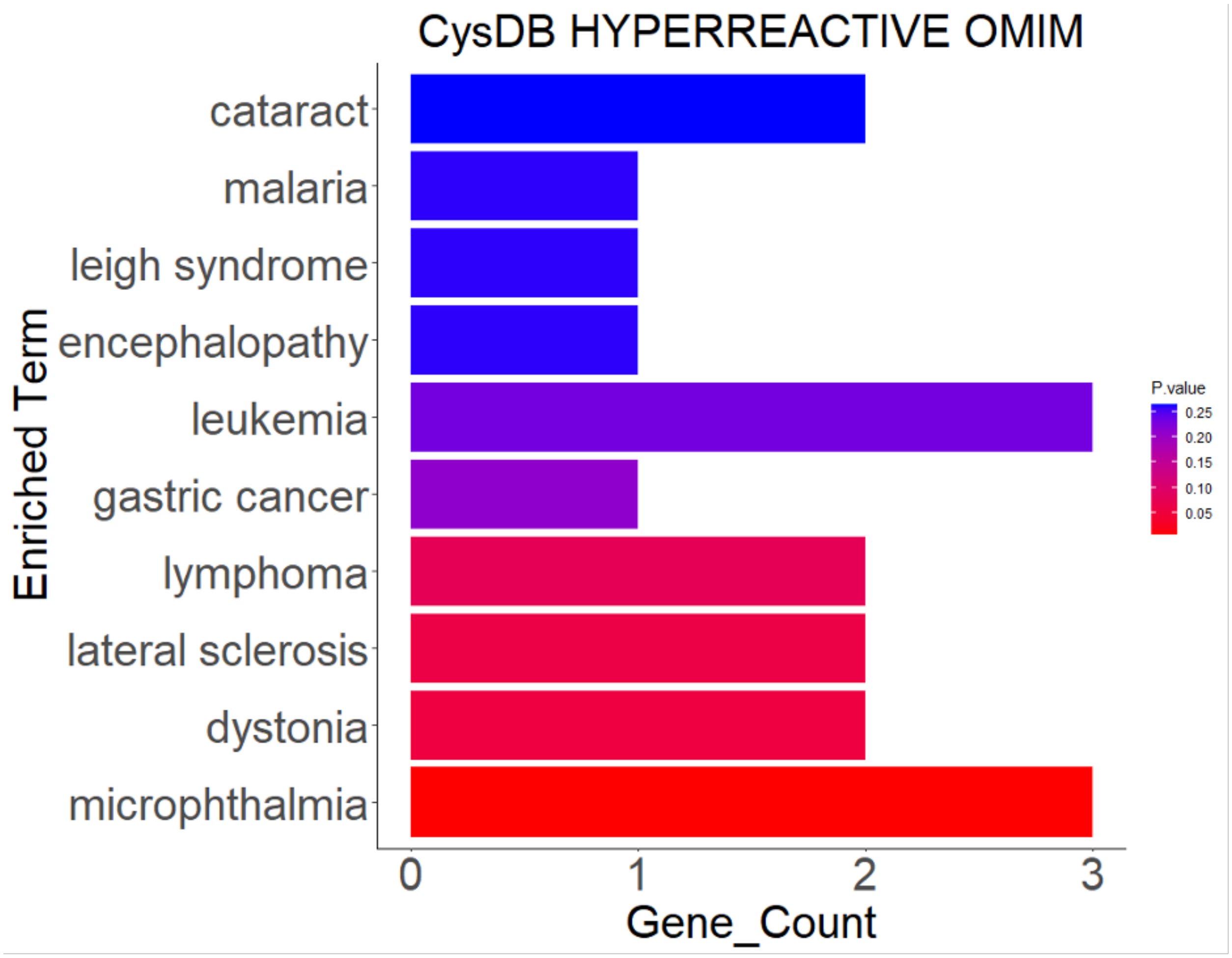


Figure S27. Top 10 enriched pathways from OMIM-term enrichment analysis of hyperreactive proteins with gene counts, related to **Figure 6**.

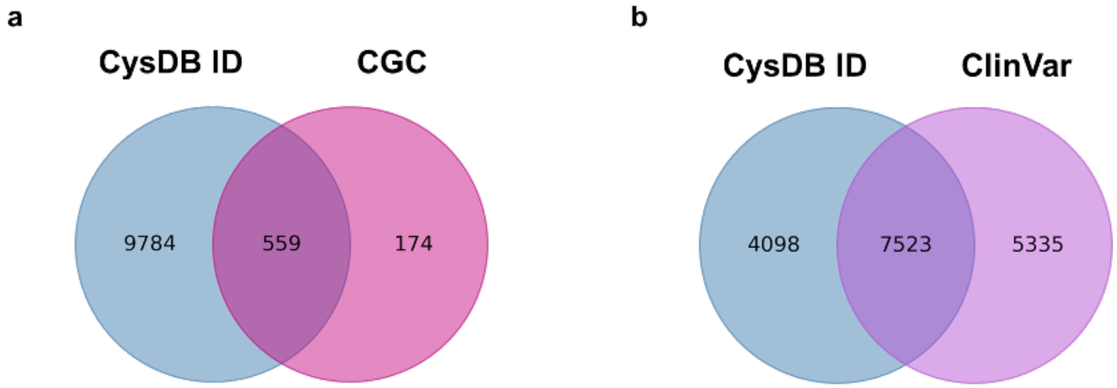


Figure S28. Overlap between the number of genes associated with CysDB identified proteins and Cancer Gene Census (CGC) genes (a), related to **Figure 6**. Overlap between the number of CysDB identified proteins and proteins associated with ClinVar variants (b).

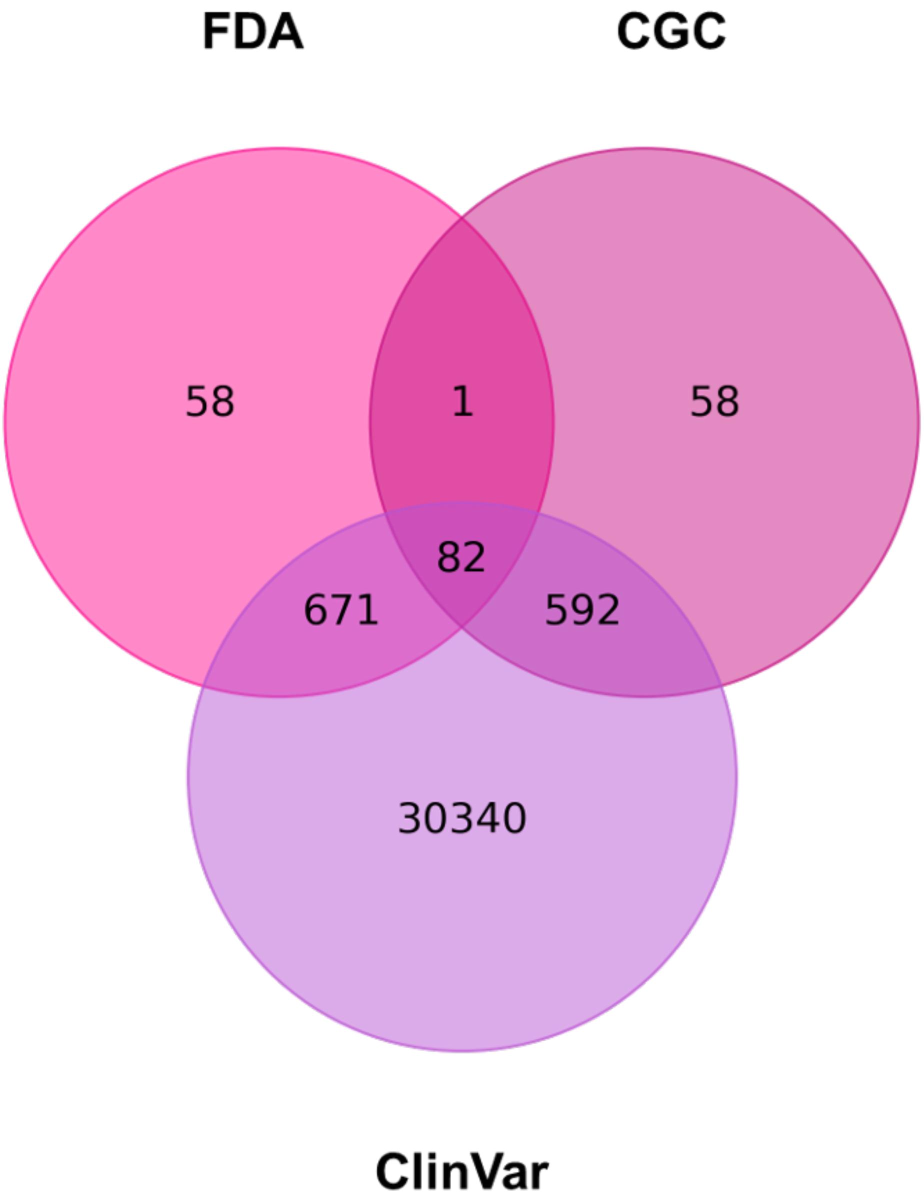


Figure S29. Overlap between the number of FDA targeted genes, Cancer Gene Census (CGC) genes and genes associated with ClinVar variants, related to **Figure 6**.

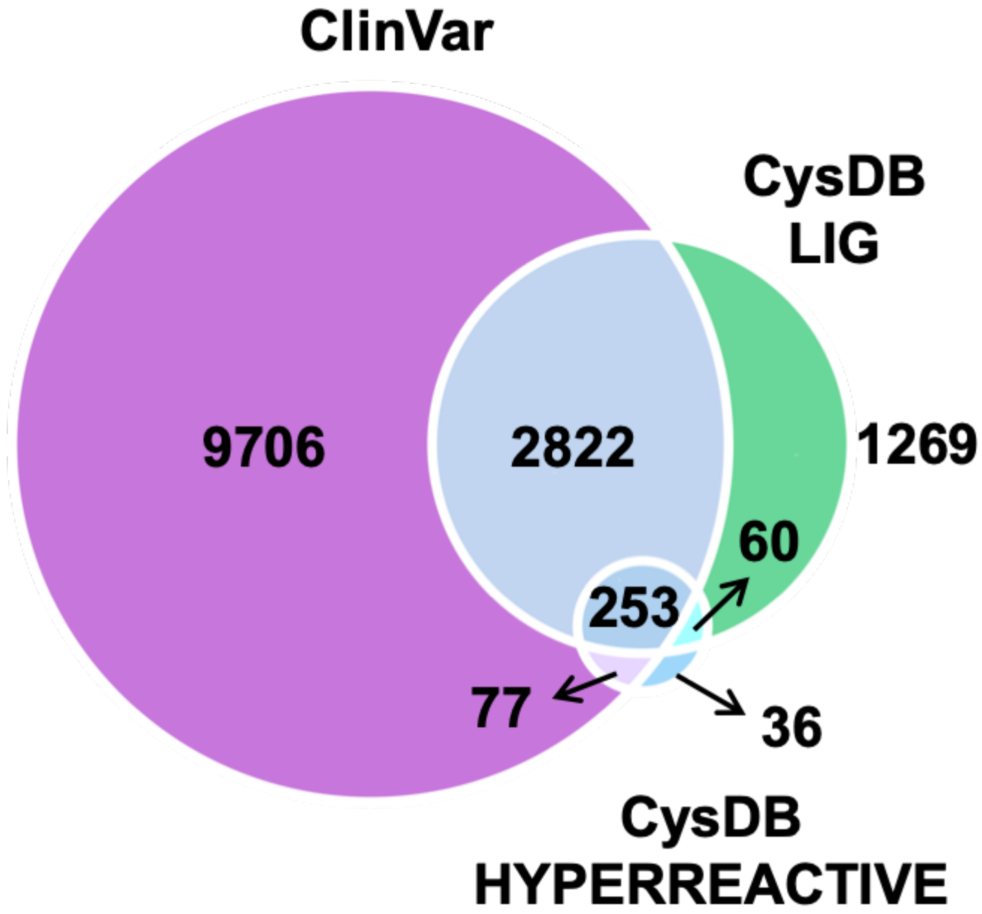


Figure S30. Overlap between the number of CysDB LIG, CysDB HYPERREACTIVE proteins and proteins associated with ClinVar variants, related to **Figure 6**.

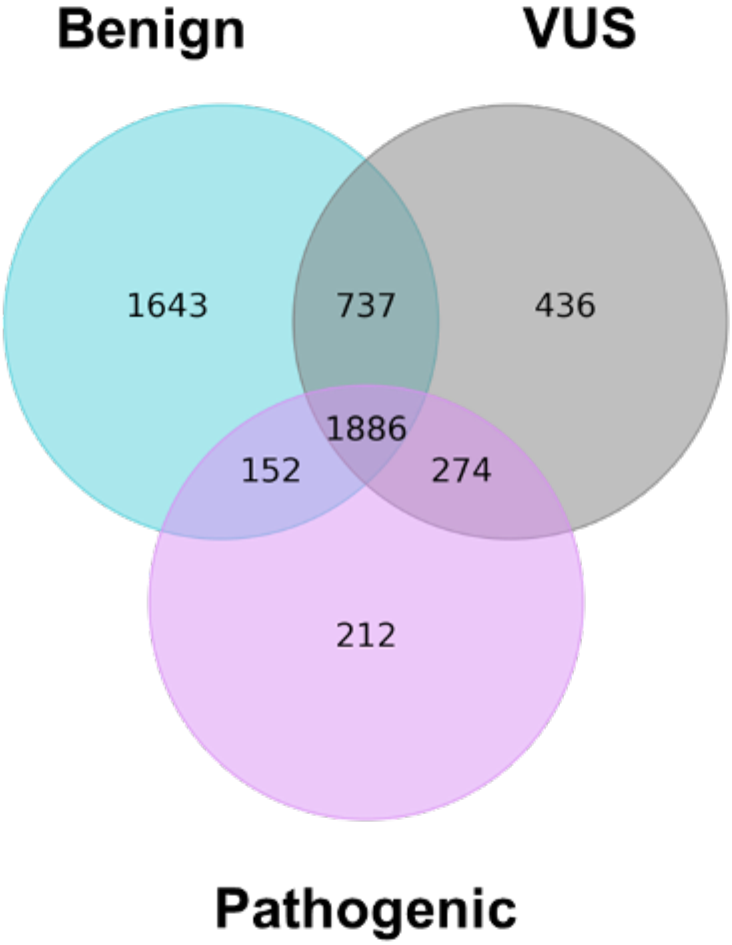


Figure S31. Overlap between the number of benign, variants of unknown significance (VUS) and pathogenic ClinVar missense variants for CysDB ID proteins, related to **Figure 6**.

2.6 - References

1. Xiao, H., Jedrychowski, M.P., Schweppe, D.K., Huttlin, E.L., Yu, Q., Heppner, D.E., Li, J., Long, J., Mills, E.L., Szpyt, J., et al. (2020). A quantitative tissue-specific landscape of protein redox regulation during aging. Cell *180*, 968–983.e24.
2. Kuljanin, M., Mitchell, D.C., Schweppe, D.K., Gikandi, A.S., Nusinow, D.P., Bulloch, N.J., Vinogradova, E.V., Wilson, D.L., Kool, E.T., Mancias, J.D., et al. (2021). Reimagining high-throughput profiling of reactive cysteines for cell-based screening of large electrophile libraries. Nat. Biotechnol. *39*, 630–641.
3. Muller, S., Ackloo, S., Al Chawaf, A., Al-Lazikani, B., Antolin, A., Baell, J.B., € Beck, H., Beedie, S., Betz, U.A.K., Bezerra, G.A., et al. (2022). Target 2035–update on the quest for a probe for every protein. RSC Med. Chem. *13*, 13–21.
4. 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.
5. Cao, J., Boatner, L.M., Desai, H.S., Burton, N.R., Armenta, E., Chan, N.J., Castello´ n, J.O., and Backus, K.M. (2021). Multiplexed CuAAC Suzuki Miyaura labeling for tandem activity-based chemoproteomic profiling. Anal. Chem. *93*, 2610–2618. https://doi.org/10.1021/acs.analchem. 0c04726.
6. Li, Z., Liu, K., Xu, P., and Yang, J. (2022). Benchmarking cleavable biotin tags for peptide-centric chemoproteomics. J. Proteome Res. *21*, 1349– 1358. https://doi.org/10.1021/acs.jproteome.2c00174.
7. Weerapana, E., Wang, C., Simon, G.M., Richter, F., Khare, S., Dillon, M.B.D., Bachovchin, D.A., Mowen, K., Baker, D., and Cravatt, B.F. (2010). Quantitative reactivity profiling

predicts functional cysteines in proteomes. Nature *468*, 790–795. https://doi.org/10.1038/nature09472.

8. Vinogradova, E.V., Zhang, X., Remillard, D., Lazar, D.C., Suciu, R.M., Wang, Y., Bianco, G., Yamashita, Y., Crowley, V.M., Schafroth, M.A., et al. (2020). An activity-guided map of electrophile-cysteine interactions in primary human T cells. Cell *182*, 1009–1026.e29.
9. 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. https:// doi.org/10.15252/msb.20209840.
10. Yang, F., Jia, G., Guo, J., Liu, Y., and Wang, C. (2022). Quantitative chemo proteomic profiling with data-independent acquisition-based mass spectrometry. J. Am. Chem. Soc. *144*, 901–911. https://doi.org/10.1021/jacs. 1c11053.
11. Backus, K.M., Correia, B.E., Lum, K.M., Forli, S., Horning, B.D., Gonzalez Paez, 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.
12. 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.
13. Backus, K.M. (2018). Applications of Reactive Cysteine Profiling (Activity Based Protein Profiling).
14. Abegg, D., Frei, R., Cerato, L., Prasad Hari, D., Wang, C., Waser, J., and Adibekian, A. (2015). Proteome-wide profiling of targets of cysteine reactive small molecules by using ethynyl benziodoxolone reagents. Angew. Chem. *54*, 10852–10857.
15. Kulkarni, R.A., Bak, D.W., Wei, D., Bergholtz, S.E., Briney, C.A., Shrimp, J.H., Alpsoy, A.,

Thorpe, A.L., Bavari, A.E., Crooks, D.R., et al. (2019). A chemoproteomic portrait of the oncometabolite fumarate. Nat. Chem. Biol. *15*, 391–400.

16. Grossman, E.A., Ward, C.C., Spradlin, J.N., Bateman, L.A., Huffman, T.R., Miyamoto, D.K., Kleinman, J.I., and Nomura, D.K. (2017). Covalent ligand discovery against druggable hotspots targeted by anti-cancer natural products. Cell Chem. Biol. *24*, 1368–1376.e4.
17. Tian, C., Sun, R., Liu, K., Fu, L., Liu, X., Zhou, W., Yang, Y., and Yang, J. (2017). Multiplexed thiol reactivity profiling for target discovery of electrophilic natural products. Cell Chem. Biol. *24*, 1416–1427.e5.
18. Wang, C., Weerapana, E., Blewett, M.M., and Cravatt, B.F. (2014). A chemoproteomic platform to quantitatively map targets of lipid-derived electrophiles. Nat. Methods *11*, 79–85.
19. 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.
20. Fu, L., Li, Z., Liu, K., Tian, C., He, J., He, J., He, F., Xu, P., and Yang, J. (2020). A quantitative thiol reactivity profiling platform to analyze redox and electrophile reactive cysteine proteomes. Nat. Protoc. *15*, 2891– 2919. https://doi.org/10.1038/s41596-020-0352-2.
21. Desai, H.S., Yan, T., Yu, F., Sun, A.W., Villanueva, M., Nesvizhskii, A.I., and Backus, K.M. (2022). SP3-Enabled rapid and high coverage chemoproteomic identification of cell-state–dependent redox-sensitive cysteines. Mol. Cell. Proteomics *21*, 100218.
22. Shi, Y., Fu, L., Yang, J., and Carroll, K.S. (2021). Wittig reagents for chemo selective sulfenic acid ligation enables global site stoichiometry analysis and redox-controlled mitochondrial

targeting. Nat. Chem. *13*, 1140–1150.

23. Mnatsakanyan, R., Markoutsa, S., Walbrunn, K., Roos, A., Verhelst, S.H.L., and Zahedi, R.P. (2019). Proteome-wide detection of S-nitrosylation targets and motifs using bioorthogonal cleavable-linker based enrichment and switch technique. Nat. Commun. *10*, 2195.
24. Wu, S., Luo Howard, H., Wang, H., Zhao, W., Hu, Q., and Yang, Y. (2016). Cysteinome: the first comprehensive database for proteins with targetable cysteine and their covalent inhibitors. Biochem. Biophys. Res. Commun. *478*, 1268–1273.
25. Yan, T., Palmer, A.B., Geiszler, D.J., Polasky, D.A., Boatner, L.M., Burton, N.R., Armenta, E., Nesvizhskii, A.I., and Backus, K.M. (2022). Enhancing cysteine chemoproteomic coverage through systematic assessment of click chemistry product Fragmentation. Anal. Chem. *94*, 3800–3810. https://doi.org/10.1021/acs.analchem.1c04402.
26. UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. *47*, 506–515.
27. Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I., and Forbes, S.A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat. Rev. Cancer *18*, 696–705.
28. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. *46*, 1062–1067.
29. Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., et al. (2010). Towards a knowledge-based human protein atlas. Nat. Biotechnol. *28*, 1248–1250.
30. Mendez, D., Gaulton, A., Bento, A.P., Chambers, J., De Veij, M., Fe´ lix, E., Magarin˜ os,

M.P., Mosquera, J.F., Mutowo, P., Nowotka, M., et al. (2019). ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res. *47*, 930–940.

31. Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. *46*, 1074–1082.
32. Rose, P.W., Prlic, A., Altunkaya, A., Bi, C., Bradley, A.R., Christie, C.H., and Burley, S.K. (2016). The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Res. gkw1000.
33. Eng, J.K., McCormack, A.L., and Yates, J.R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. *5*, 976–989. https://doi. org/10.1016/1044-0305(94)80016-2.
34. Yu, F., Teo, G.C., Kong, A.T., Haynes, S.E., Avtonomov, D.M., Geiszler, D.J., and Nesvizhskii, A.I. (2020). Identification of modified peptides using localization-aware open search. Nat. Commun. *11*, 4065. https://doi.org/ 10.1038/s41467-020-17921-y.
35. Integrated Proteomics Pipeline (IP2). http://www.integratedproteomics. com/
36. Xu, T., Park, S.K., Venable, J.D., Wohlschlegel, J.A., Diedrich, J.K., Cociorva, D., Lu, B., Liao, L., Hewel, J., Han, X., et al. (2015). ProLuCID: an improved SEQUEST-like algorithm with enhanced sensitivity and specificity. J. Proteomics *129*, 16–24. https://doi.org/10.1016/j.jprot.2015. 07.001.
37. Kong, A.T., Leprevost, F.V., Avtonomov, D.M., Mellacheruvu, D., and Nesvizhskii, A.I. (2017). MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. Nat. Methods *14*, 513–520. https://doi.org/10.1038/nmeth.4256.

38. Eng, J.K., Jahan, T.A., and Hoopmann, M.R. (2013). Comet: an open source MS/MS sequence database search tool. Proteomics *13*, 22–24. https://doi.org/10.1002/pmic.201200439.

39. Serafimova, I.M., Pufall, M.A., Krishnan, S., Duda, K., Cohen, M.S., Maglathlin, R.L., McFarland, J.M., Miller, R.M., Fro¨ din, M., and Taunton, J. (2012). Reversible targeting of noncatalytic cysteines with chemically tuned electrophiles. Nat. Chem. Biol. *8*, 471–476.

40. 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.

41. Abbasov, M.E., Kavanagh, M.E., Ichu, T.A., Lazear, M.R., Tao, Y., Crowley, V.M., Am Ende, C.W., Hacker, S.M., Ho, J., Dix, M.M., et al. (2021). A proteome-wide atlas of lysine-reactive chemistry. Nat. Chem. *13*, 1081–1092.

42. Braschi, B., Denny, P., Gray, K., Jones, T., Seal, R., Tweedie, S., Yates, B., and Bruford, E. (2019). Genenames. org: the HGNC and VGNC resources in 2019. Nucleic Acids Res. *47*, 786–792.

43. The Gene Ontology Consortium (2019). The gene ontology resource: 20 years and still GOing strong. Nucleic Acids Res. *47*, 330–338.

44. Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., D'Eustachio, P., Jassal, B., Korninger, F., May, B., et al. (2018). The reactome pathway knowledgebase. Nucleic Acids Res. *46*, 649–655.

45. Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinf. *14*, 128.

46. Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. *44*, W90–W97.

47. Schoenmaker, L., Bequignon, O.J., Jespers, W., and Westen, G.J. (2023). UnCorrupt SMILES: a novel approach to de novo design. Journal of Cheminformatics *15*, 22.

48. Bickerton, G.R., Paolini, G.V., Besnard, J., Muresan, S., and Hopkins, A.L. (2012). Quantifying the chemical beauty of drugs. Nat. Chem. *4*, 90–98.

49. Benet, L.Z., Hosey, C.M., Ursu, O., and Oprea, T.I. (2016). BDDCS, the Rule of 5 and druggability. Adv. Drug Deliv. Rev. *101*, 89–98.

50. Lipinski, C.A., Lombardo, F., Dominy, B.W., and Feeney, P.J. (2012). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug Deliv. Rev. *64*, 4–17.

51. Ghose, A.K., Viswanadhan, V.N., and Wendoloski, J.J. (1999). A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. J. Comb. Chem. *1*, 55–68.

52. Congreve, M., Carr, R., Murray, C., and Jhoti, H. (2003). A'rule of three'for fragment-based lead discovery? Drug Discov. Today *8*, 876–877.

53. Landrum, G. (2013). Rdkit documentation. Release *1*, 1–79.

54. Senkane, K., Vinogradova, E.V., Suciu, R.M., Crowley, V.M., Zaro, B.W., Bradshaw, J.M., Brameld, K.A., and Cravatt, B.F. (2019). The proteome-wide potential for reversible covalency at cysteine. Angew. Chem. *58*, 11385–11389.

55. Krishnan, S., Miller, R.M., Tian, B., Mullins, R.D., Jacobson, M.P., and Taunton, J. (2014).

Design of reversible, cysteine-targeted Michael acceptors guided by kinetic and computational analysis. J. Am. Chem. Soc. *136*, 12624–12630.

56. Zambaldo, C., Vinogradova, E.V., Qi, X., Iaconelli, J., Suciu, R.M., Koh, M., Senkane, K., Chadwick, S.R., Sanchez, B.B., Chen, J.S., et al. (2020). 2- Sulfonylpyridines as tunable, cysteine-reactive electrophiles. J. Am. Chem. Soc. *142*, 8972–8979.

57. Du, X., Guo, C., Hansell, E., Doyle, P.S., Caffrey, C.R., Holler, T.P., McKerrow, J.H., and Cohen, F.E. (2002). Synthesis and structure activity relationship study of potent trypanocidal thiosemicarbazone inhibitors of the trypanosomal cysteine protease cruzain. J. Med. Chem. *45*, 2695–2707.

58. Greenbaum, D.C., Mackey, Z., Hansell, E., Doyle, P., Gut, J., Caffrey, C.R., Lehrman, J., Rosenthal, P.J., McKerrow, J.H., and Chibale, K. (2004). Synthesis and structure activity relationships of parasiticidal thiosemicarbazone cysteine protease inhibitors against Plasmodium falciparum, Trypanosoma brucei, and Trypanosoma cruzi. J. Med. Chem. *47*, 3212–3219.

59. Shenai, B.R., Lee, B.J., Alvarez-Hernandez, A., Chong, P.Y., Emal, C.D., Neitz, R.J., Roush, W.R., and Rosenthal, P.J. (2003). Structure-activity relationships for inhibition of cysteine protease activity and development of Plasmodium falciparum by peptidyl vinyl sulfones. Antimicrob. Agents Chemother. *47*, 154–160.

60. Kluver, E., Schulz-Maronde, S., Scheid, S., Meyer, B., Forssmann, W.G., € and Adermann, K. (2005). Structure activity relation of human b-defensin 3: influence of disulfide bonds and cysteine substitution on antimicrobial activity and cytotoxicity. Biochemistry *44*, 9804–9816.

61. Grzonka, Z., Jankowska, E., Kasprzykowski, F., Kasprzykowska, R., Lankiewicz, L., Wiczk, W., Wieczerzak, E., Ciarkowski, J., Drabik, P., Janowski, R., et al. (2001). Structural studies

of cysteine proteases and their inhibitors. Acta Biochim. Pol. *48*, 1–20.

62. Zanon, P.R., Yu, F., Musacchio, P., Lewald, L., Zollo, M., Krauskopf, K., and Hacker, S.M. (2021). Profiling the Proteome-wide Selectivity of Diverse Electrophiles. ChemRxiv. https://doi.org/10.26434/chemrxiv.14186561.v1.

63. Dana, J.M., Gutmanas, A., Tyagi, N., Qi, G., O'Donovan, C., Martin, M., and Velankar, S. (2019). SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. Nucleic Acids Res. *47*, 482–489.

64. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., et al. (2021). Pfam: the protein families database in 2021. Nucleic Acids Res. *49*, 412–419.

65. Julio, A.R., and Backus, K.M. (2021). New approaches to target RNA binding proteins. Curr. Opin. Chem. Biol. *62*, 13–23.

66. de la Cruz, J., Kressler, D., and Linder, P. (1999). Unwinding RNA in Saccharomyces cerevisiae: DEAD-box proteins and related families. Trends Biochem. Sci. *24*, 192–198.

67. Aubourg, S., Kreis, M., and Lecharny, A. (1999). The DEAD box RNA helicase family in Arabidopsis thaliana. Nucleic Acids Res. *27*, 628–636.

68. Patmore, D.M., Jassim, A., Nathan, E., Tong, Y., Tahan, D., Hoffmann, N., Gilbertson, R.J., Smith, K.S., Kanneganti, T.D., Suzuki, H., et al. (2020). DDX3X suppresses the susceptibility of hindbrain lineages to medulloblastoma. Dev. Cell *54*, 455–470.e5.

69. Andrisani, O., Liu, Q., Kehn, P., Leitner, W.W., Moon, K., Vazquez Maldonado, N., and Gale, M. (2022). Biological Functions of DEAD/ DEAH-box RNA Helicases in Health and Disease. Nature Immunology *23*, 354–357.

70. Mi, H., Muruganujan, A., Ebert, D., Huang, X., and Thomas, P.D. (2019). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res. *47*, 419–426.

71. Fesik, S.W. (2005). Promoting apoptosis as a strategy for cancer drug discovery. Nat. Rev. Cancer *5*, 876–885.

72. Aguilar, A., Lu, J., Liu, L., Du, D., Bernard, D., McEachern, D., Przybranowski, S., Li, X., Luo, R., Wen, B., et al. (2017). Discovery of 4-((3′ R, 4′ S, 5′ R)-6 ″-Chloro-4′-(3-chloro-2-fluorophenyl)-1′-ethyl-2 ″-oxodispiro [cyclohexane-1, 2′-pyrrolidine-3′, 3 ″-indoline]-5′-carboxamido) bicyclo [2.2.2] octane-1-carboxylic acid (AA-115/APG-115): a potent and orally active murine double minute 2 (MDM2) inhibitor in clinical development. J. Med. Chem. *60*, 2819–2839.

73. Giancotti, F.G., and Ruoslahti, E. (1999). Integrin signaling. Science *285*, 1028–1032.

74. Cooper, J., and Giancotti, F.G. (2019). Integrin signaling in cancer: mechanotransduction, stemness, epithelial plasticity, and therapeutic resistance. Cancer Cell *35*, 347–367.

75. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., and McKusick, V.A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledge base of human genes and genetic disorders. Nucleic Acids Res. *33*, D514–D517. https://doi.org/10.1093/nar/gki033.

76. Schrijver, I., Liu, W., Brenn, T., Furthmayr, H., and Francke, U. (1999). Cysteine substitutions in epidermal growth factor–like domains of fibrillin-1: distinct effects on biochemical and clinical phenotypes. Am. J. Hum. Genet. *65*, 1007–1020.

77. Russell, D.W., Brown, M.S., and Goldstein, J.L. (1989). Different combinations of cysteine-rich repeats mediate binding of low density lipoprotein receptor to two different proteins. J. Biol. Chem. *264*, 21682–21688.

78. Daly, N.L., Scanlon, M.J., Djordjevic, J.T., Kroon, P.A., and Smith, R. (1995). Three-dimensional structure of a cysteine-rich repeat from the low-density lipoprotein receptor. Proc. Natl. Acad. Sci. USA *92*, 6334–6338.

79. Esser, V., Limbird, L.E., Brown, M.S., Goldstein, J.L., and Russell, D.W. (1988). Mutational analysis of the ligand binding domain of the low density lipoprotein receptor. J. Biol. Chem. *263*, 13282–13290.

80. Lanman, B.A., Allen, J.R., Allen, J.G., Amegadzie, A.K., Ashton, K.S., Booker, S.K., and Cee, V.J. (2020). Discovery of a Covalent Inhibitor of KRASG12C (AMG 510) for the Treatment of Solid Tumors. J. Med. Chem. *63*, 52–65.

81. Janes, M.R., Zhang, J., Li, L.S., Hansen, R., Peters, U., Guo, X., Chen, Y., Babbar, A., Firdaus, S.J., Darjania, L., et al. (2018). Targeting KRAS mutant cancers with a covalent G12C-specific inhibitor. Cell *172*, 578–589.e17.

82. 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 StateTargeting inactive KRASG12C suppresses oncogenic signaling. Cancer Discov. *6*, 316–329.

83. 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.

84. Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M.Y., Geiger, T., Mann, M., and Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote) omics data. Nat. Methods *13*, 731–740.

85. Perez-Riverol, Y., Bai, J., Bandla, C., Garcı´a-Seisdedos, D., Hewapathirana, S., Kamatchinathan, S., Vizcaı´no, J.A., Prakash, A., Frericks-Zipper, A., Eisenacher, M., et al.

(2022). The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. Nucleic Acids Res. *50*, 543–552.

86. Han, X., Wang, J., Wang, J., Liu, S., Hu, J., Zhu, H., and Qian, J. (2017). ScaPD: a database for human scaffold proteins. BMC Bioinf. *18*, 386.

87. Xie, Z., Bailey, A., Kuleshov, M.V., Clarke, D.J.B., Evangelista, J.E., Jenkins, S.L., Lachmann, A., Wojciechowicz, M.L., Kropiwnicki, E., Jagodnik, K.M., et al. (2021). Gene set knowledge discovery with enrichr. Curr. Protoc. *1*, e90.

Chapter 3

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], Peiyuan Lee[4], Kendall N. Houk[2], Stefano Forli[3]* and Keriann M. Backus[1,2,5]*

[1]Biological Chemistry Department, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA, [2]Department of Chemistry and Biochemistry, University of California, Los Angeles, Los Angeles, CA 90095, USA, [3]Department of Integrative Structural and Computational Biology, Scripps Research, La Jolla, CA, 92037, USA, [4]Department of Statistics and Data Science, UCLA, Los Angeles, CA, 90095, USA, [5]Lead contact *Correspondence: forli@scripps.edu or kbackus@mednet.ucla.edu

3.1 - 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-reactive 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-reactive cysteines. To build CIAA we first generated a high coverage proteomic dataset of high-reactive cysteines that features 823 total high-reactive cysteines, identified in both newly generated and previously published datasets. We achieve >50% increase in total high-reactive cysteines when compared to prior datasets. We then subject a class-balanced set of high- and low-reactive 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.

3.2 - Results

3.2.1 - Establishing a high coverage dataset of high-reactive cysteines towards iodoacetamide alkyne (IAA)

Our first step towards enhancing the *in silico* discovery of high-reactive 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. 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] In these studies, 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.

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-reactive cysteines across datasets. In total these prior datasets contained 489 total high-reactive 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 this 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. In total, the relative reactivity of 9,783 cysteines from 3,974 proteins were quantified across both newly generated and previously reported data. 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**).

3.2.2 - Cysteine reactivity correlates with UniProtKB indications of functionality

As our newly generated data has more than doubled the total number of high-reactive 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-reactive cysteines were maintained in our new and larger dataset. Notably, and corroborating prior findings[6], we observe that cysteine high-reactive 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-reactive cysteines were identified with UniProtKB annotations related to zinc binding or zinc finger regions, compared to 20 high-reactive cysteines. This analysis confirmed that our newly generated data did extend cysteine coverage while showing a similar properties distribution of previously reported datasets.

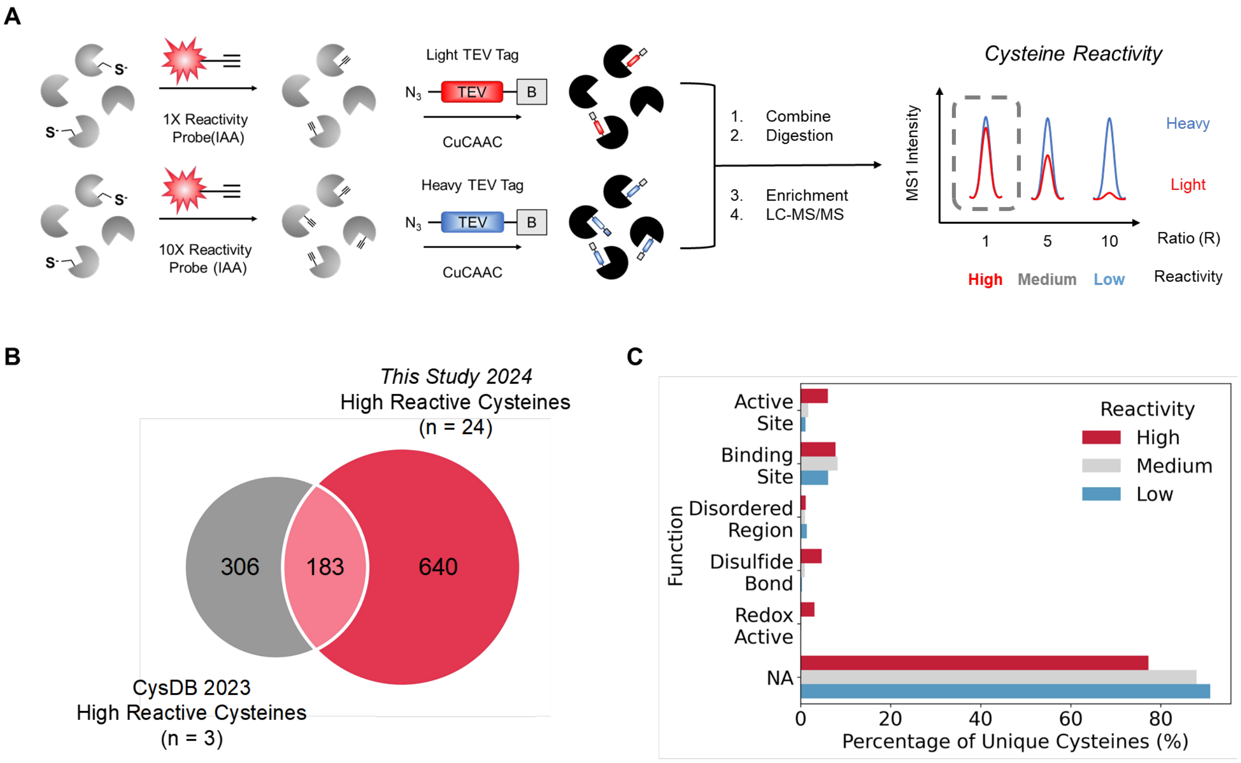


Figure 1. Establishing a high coverage dataset of high-reactive cysteines towards iodoacetamide alkyne (IAA). (A) Experimental workflow for isoTOP-ABPP. Cell lysates are treated with either high (100 µM) or low (10 µM) concentration of this IAA probe followed by click conjugation to isotopically differentiated tobacco etch virus (TEV)-cleavable biotinylated enrichment tags. After single pot solid phase 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-reactive 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-reactive 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-reactive cysteines. See also **Figure S1**, **Figure S2**, and **Data S1**.

3.2.3 - Primary sequence amino acid composition of high-reactive cysteines

Previous analysis of a focused set (n = 74) of high-reactive cysteines had revealed enrichment for tryptophan, histidine, proline, and cysteine residues in linear sequence proximity to high-reactive 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 subject 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-reactive cysteines, starting with 823 high-reactive cysteines and 996 low-reactive 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-reactive cysteines as the foreground and 805 low-reactive cysteines as the background (**Figure 2B** and **Data S1**). This analysis revealed an increased occurrence of cysteines (C) near high-reactive 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-reactive 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.

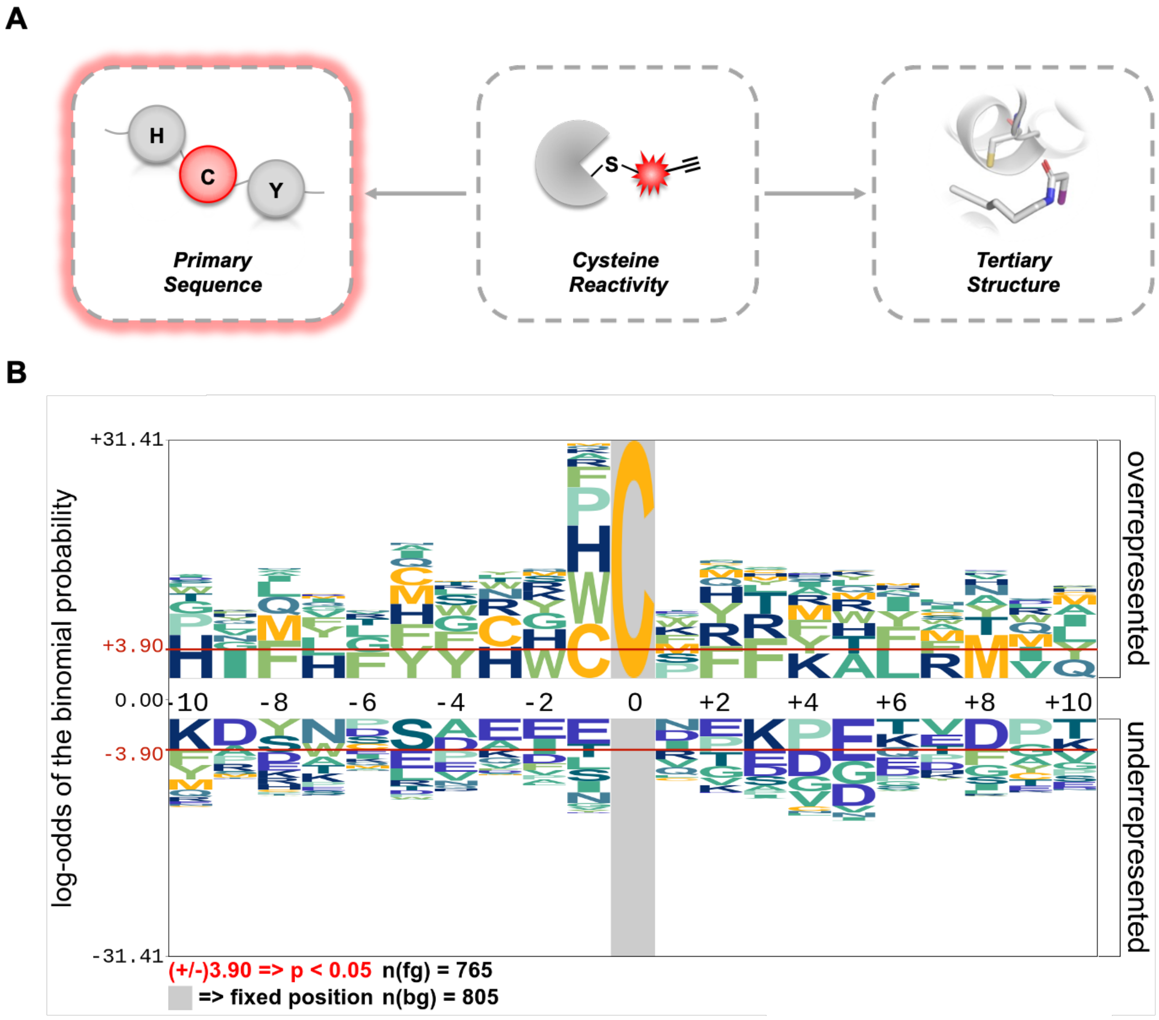


Figure 2. Amino acid contents of IAA-reactive cysteines using primary sequences. (A) Schematic highlighting Figure 2 analysis focuses on using primary sequences of IAA-labeled cysteines. (B) Sequence logo created using pLogo (http://plogo.uconn.edu).[35] Starting with 823 high-reactive and 996 low-reactive cysteines, sequences were aligned to meet pLogo input requirements, reducing the dataset to 765 high-reactive cysteines as the foreground and 805 low-reactive 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**.

3.2.4 - 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-reactive 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 PDBs and those experimentally labeled by IAA (**Figure S4**). However, we observed an enrichment of enzyme structures among proteins with PDBs, 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-reactive cysteines and 322 out of 996 (33%) low-reactive 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-reactive and 297 low-reactive cysteines across 644 and 662 unique PDBs (**Figure 3C** and **Figure S5A**). Notably, 32 of these proteins contained both a high-reactive and a low-reactive cysteine. Importantly, and demonstrating that our filtering steps did not introduce significant bias to the datasets, the high-reactive and low-reactive protein sets exhibited similar distributions of experimental techniques, structural resolutions, and biological complexes (**Figure S6**).

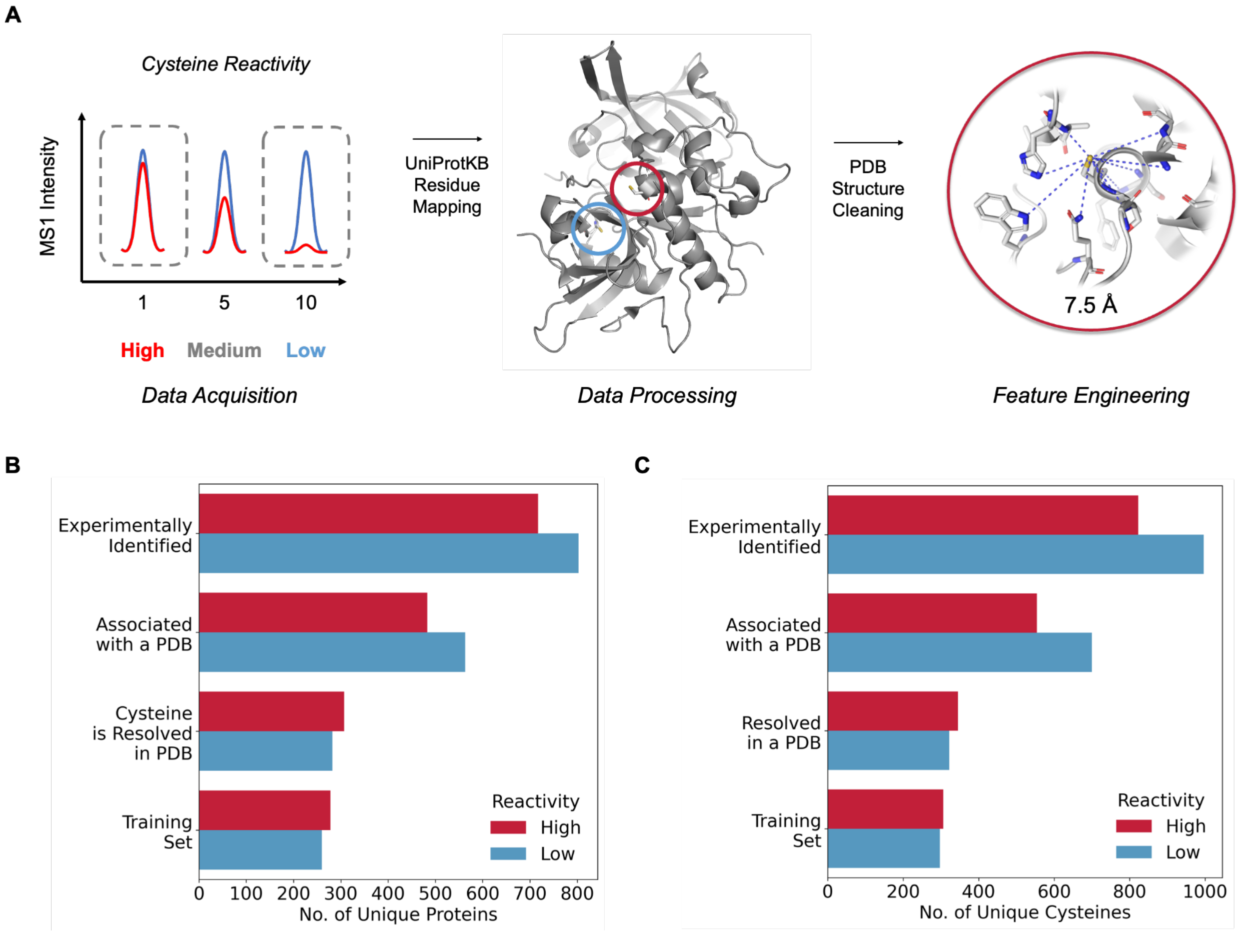


Figure 3. Defining a training set of reactive cysteines with 3D structural data available in the PDB. (A) Workflow for defining a training set of tertiary structures. (B) Bar graph showing the number of experimentally identified proteins containing high- or low-reactive cysteines, number of experimentally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

associated PDB structure, and number of experimentally identified unique high- or low-reactive cysteines in the training set after a series of filtering steps. See **Figure S3-S6**, and **Data S2**.

3.2.5 - Tertiary structure amino acid composition of hyper-reactive cysteines

With our curated set of structurally resolved cysteines in hand, we next sought to assess the amino acid content of IAA high-reactive cysteine 3D neighborhoods (**Figure 4A**). Similar to our linear sequence analysis (**Figure 2**), we hypothesized that the 3D protein environment surrounding high-reactive 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 reactive and unreactive 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 the 7.5 Å distance as it provided a balanced approach to capturing neighboring residues without sampling more distal residues (**Figure S7**). This threshold was chosen after testing 5, 7.5, and 10 Å cutoffs. The 5 Å radius resulted in a higher enrichment of residues, while the 10 Å radius yielded too few, potentially missing relevant neighbors. The 7.5 Å distance offered a middle ground, capturing an appropriate number of residues without over- or under-representation.

To prevent overcounting and 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-reactive 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-reactive 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-reactive cysteine-specific amino acid enrichment represents bona fide features within the 3D cysteine microenvironment that drive cysteine nucleophilicity.

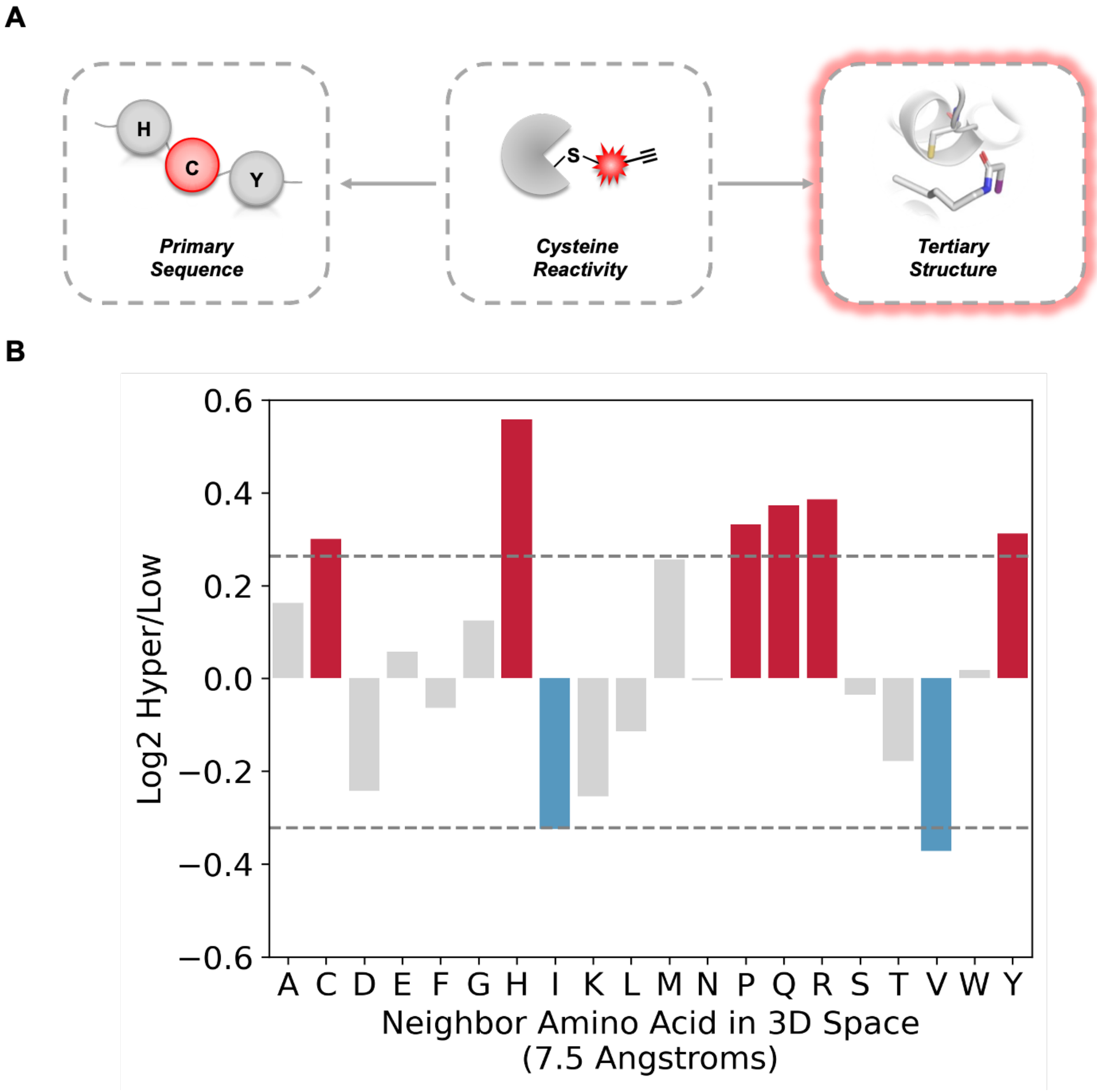


Figure 4. Amino acid content of IAA-reactive cysteines using 3D protein structures. (A) Schematic highlighting **Figure 4** analysis focuses on using tertiary structures of IAA-labeled cysteines resolved in associated PDB structures (306 high-reactive cysteines and 297 low-reactive cysteines). (B) Log2 ratio of amino acid frequencies within a 7.5 Å neighborhood around high-reactive cysteines relative to low-reactive cysteines. Red bars indicate enriched residues in high-reactive cysteine neighborhoods, while blue bars indicate depleted residues in these

neighborhoods. See **Figure S7**, **Figure S8**, and **Data S3**.

3.2.6 - Descriptors of IAA-reactive cysteines from 3D structures

As pKa prediction was insufficient to predict cysteine reactivity and amino acid content analysis had pointed towards the likelihood of clear differences between the microenvironment of high- and low-reactive cysteines, we next opted to extend our analysis to consider additional features beyond amino acid content (**Figure 5A**). To capture potential structural features of 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 focused on amino acid physicochemical properties and hydrogen bond interactions. For both 1D and 3D amino acid content, we used the aforementioned 7.5 Å cutoff and incorporated all amino acid information within this region into the descriptors. 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 interactions were defined when the distance between the cysteine's 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 detailed 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.**

3.2.7 - pKa prediction is insufficient to predict cysteine reactivity

Alongside structural features, availability of computational tools that predict cysteine pKa, most notably PROPKA[18], highlights another potential opportunity for rapid discovery of reactive cysteines. Therefore, we next sought to investigate whether 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 pKa predictions and measures of IAA-cysteine reactivity, we next expanded our analysis to a larger dataset using a predictive program to estimate pKa values across multiple cysteines. Using PROPKA[18] version 3.1, we computed the predicted pKa for each cysteine-structure pair and calculated a median predicted pKa for all structures associated with a specific UniProtKB_C# identifier, and any predicted pKa greater than 14 was set to 14 for clarity. Our 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 reactive cysteines was 11.18 versus 10.71 for low-reactive 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-reactive cysteines had predicted pKa values that strongly contrasted with their measured reactivity. Exemplifying this difference, for the high-reactive 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 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 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, six of 16 cysteines 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 6 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-reactive and low-reactive cysteine subsets, respectively.

3.2.8 - Prevalence of hyper-reactive cysteines in secondary structure motifs

Previous studies have suggested that high-reactive 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-reactive cysteines were found in helices, 36 in beta sheets, and 116 in loops. In comparison, we observed 81 low-reactive 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**).

3.2.9 - Relative solvent accessibility and pocket detection of high-reactive 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-reactive 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-reactive cysteines had a median solvent accessibility of 15%, with 26% classified as "high-solvent accessible" (RSA ≥ 20), 15.7% as medium solvent accessible (10 ≤ 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-reactive 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-reactive sites within potential binding pockets away from bulk solvent.

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

Guided by the suggestive enrichments for high-reactive 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-reactive cysteines (**Figure S14**). This inverse relationship suggests that less reactive cysteines (higher isoTOP-ABPP reactivity ratios) 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 leads us to conclude that cysteine high reactivity towards IAA is likely governed by a combination of factors rather than any single structural feature.

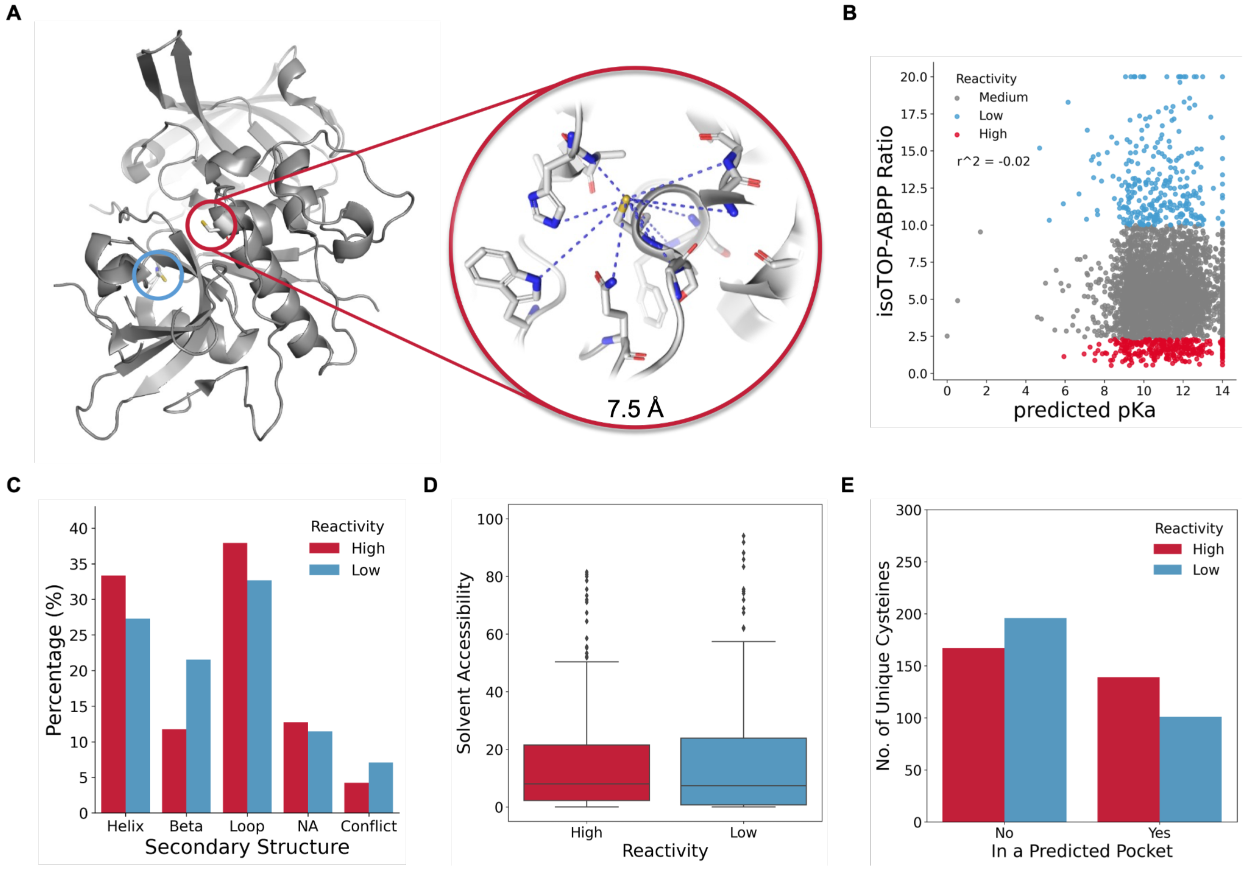


Figure 5. Chemical, reactivity and structural properties of IAA-reactive cysteines. (A) Collection of chemical, reactivity, and structural properties of IAA-reactive cysteines using tertiary structures. (B) Comparing computationally predicted pKa (PROPKA 3.1)[18] values and quantitative cysteine reactivity 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**.

3.2.11 - 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-reactive (0) or high-reactive (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-reactive cysteine class.

To maximize the number of high-reactive 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.

3.2.12 - Ablation studies for 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.

3.2.13 - Effect of multiple structural representations on prediction accuracy

For example, the CIAA method predicted C141 of Flap Endonuclease (FEN1) to be high-reactive 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-reactive. 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-reactive. 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.

3.2.14 - 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-reactive (0) or high-reactive (1) cysteines.[62,68] In our analysis, positive SHAP values indicated an increased likelihood of predicting high-reactive 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-reactive 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-reactive C100 in the Multifunctional methyltransferase subunit TRM112-like protein (TRNT112) and low-reactive 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.

3.2.15 - Model limitations and performance across protein functional classes

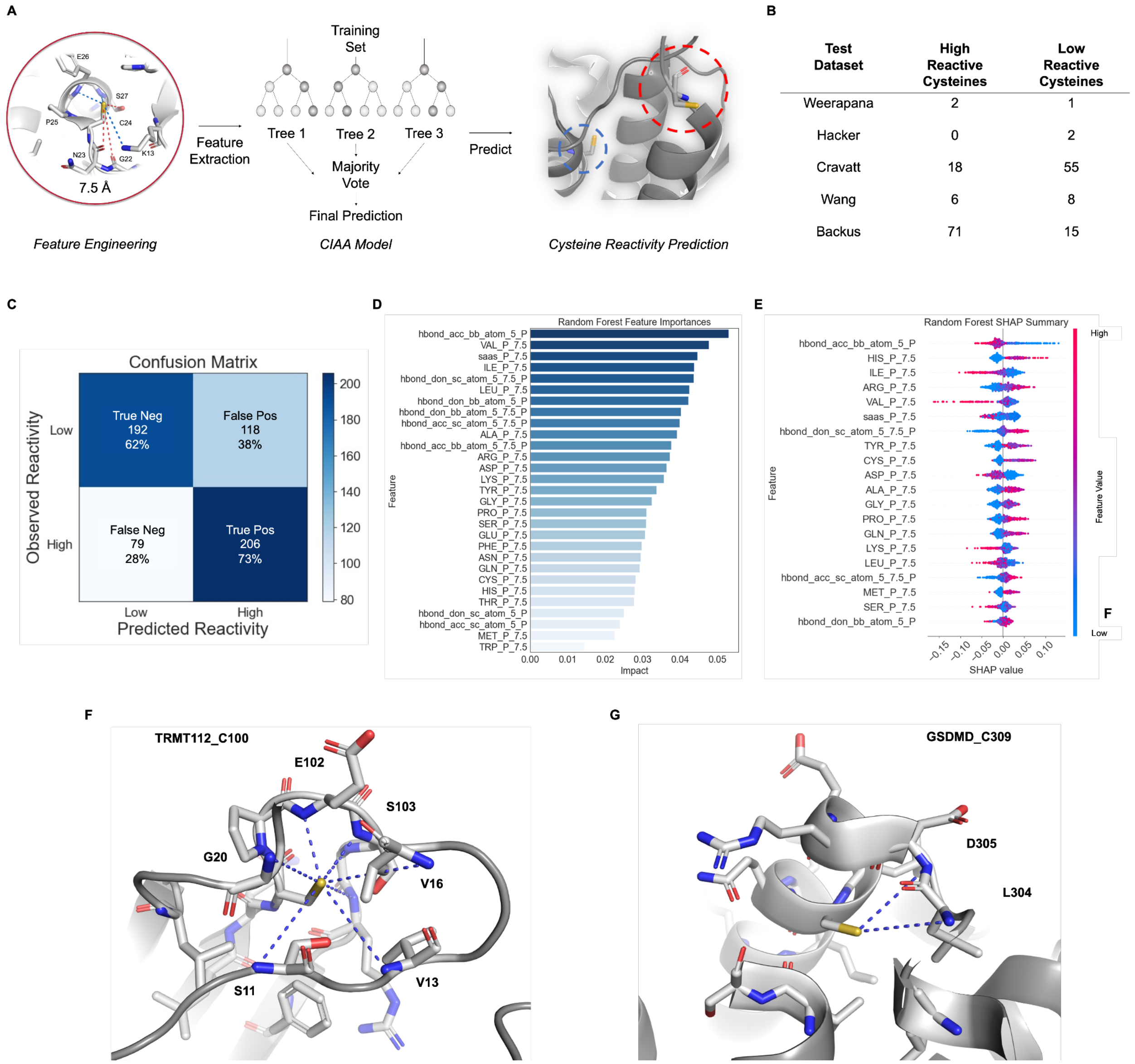
It is important to acknowledge the limitations of our model by examining cases where it failed. We compared correct and incorrect predictions across experimental structure determination methods. The model performed consistently across methods, with the highest TPR of 69% for X-ray structures (n = 370 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-reactive 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-reactive cysteines in nucleic acid/small molecule-binding proteins, chaperones, transporters, channels, and receptors (**Figure S21**). However, it struggled with correctly predicting low-reactive cysteines for enzymes, leading to an increase in false positives. Most high-reactive 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-reactive (**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-reactive C14 of Uroporphyrinogen-III synthase (UROS) ($R_{10:1}$ = 19.22) was incorrectly predicted to be high-reactive. 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).

3.2.16 - 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-reactive cysteines incorrectly predicted as low-reactive (**Figure S23B**).



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-reactive 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-reactive) 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-reactive cysteines. (E) SHapley Additive exPlanations (SHAP) summary showing the impact of selected features on the predicted classification (high- or low-reactive 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-reactive 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-reactive 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**.

3.3 - Discussion

To enhance the discovery of high-reactive 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-reactive—thus our work more than doubles the number of known high-reactive 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, find that high-reactive 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-reactive 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-reactive, was predicted to be high-reactive by the CIAA model when analyzed in the active conformation of CASP1 (PDB: 6BZ9)-we expect this disconnect stems from the nonapoptotic nature of the 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

Analyzing over 1,000 Protein Data Bank (PDB) structures alongside computationally predicted AlphaFold 2 models, CIAA distinguishes high-reactive cysteines from low-reactive ones with notable accuracy. This work represents a significant step forward in merging proteomics with structural biology, providing a robust tool for exploring cysteine reactivity towards IAA and paving the way for new applications in drug discovery and protein research. Here, our first-pass attempts at realizing this vision fell short, likely due to side chain conformations that differed from experimentally determined structures,[74] which diminished the model performance. We are optimistic that future implementations of AlphaFold 2 and related tools will prove compatible with *in silico* cysteine analysis.[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.), and NIGMS System and Integrative Biology 5T32GM008185-33 (L.M.B.). We thank all members of the Houk, Forli, and Backus labs for helpful suggestions.

Author Contributions

L.M.B., 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., and K.N.H. provided technical advice. L.M.B., S.F. and K.M.B. wrote the manuscript.

Declaration of Interests

The authors declare no financial or commercial conflict of interest.

3.4 Methods

Data Storage

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.

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 **Table 3.1**.

Code Availability

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

3.4.1 - Experimental Methods

Cell culture.

Cell culture reagents including Dulbecco's phosphate-buffered saline (DPBS), Roswell Park Memorial Institute (Gibco™ RPMI 1640 Medium, 11875119) media, Dulbecco's Modified Eagle Medium (DMEM, Gibco™ 11995073) media and penicillin/streptomycin (Pen/Strep, Gibco™, Penicillin-Streptomycin 10,000 U/mL, 15140122) were purchased from Fisher Scientific. Fetal Bovine Serum (FBS) was purchased from Avantor Seradigm (Avantor®, Seradigm, Premium Grade Fetal Bovine Serum, Cat.No. 97068-085, lot #214B17). All cell lines were obtained from ATCC and were maintained at a low passage number (< 20 passages). HEK293T (ATCC: CRL-3216) cells were cultured in DMEM supplemented with 10% FBS and 1% antibiotics (Penn/Strep, 100 U/mL). Jurkat (ATCC: TIB-152) cells were cultured in RPMI-

1640 supplemented with 10% FBS and 1% antibiotics (Penn/Strep, 100 U/mL). Media was filtered (0.22 μm) prior to use. Cells were maintained in a humidified incubator at 37 °C with 5% CO2. Cell line was tested monthly with the Mycoplasma detection kit (InvitroGen).

Cell harvesting and cell lysis.

Cells were harvested by centrifugation (4500g, 5 min, 4 °C) and washed twice with cold DPBS. Cell pellets were then lysed with sonication using an Ultrasonic Probe Sonicator at amplitude 2 for 10 pulses, 1 second pulse, 1 second off on ice and fractionated with ultracentrifugation (100,000 *g*, 1 hr, 4 °C). Supernatant was saved as the soluble fraction. Pellet was resuspended with 500 μL PBS, sonicated and saved as the membrane fraction. Protein concentrations were determined using a Bio-Rad DC protein assay kit (Cat. No. 5000112 BioRad Life Science) and the lysates were diluted to lysate diluted to the working concentrations indicated below.

Cysteine labeling.

After cells were harvested and lysed, as stated, lysate concentrations were then adjusted to 2 mg/mL. For cysteine reactivity quantification, lysates were then labeled with either 10 μM or 100 μM iodoacetamide alkyne (**IAA**) for 1h at room temperature (RT). Samples were then subjected to bioorthogonal copper(I)-catalyzed azide-alkyne cycloaddition (CuAAC) to conjugate isotopically labeled *tobacco etch virus* (TEV)-cleavable biotinylated peptide tags. To 200 μL cell lysates (2 mg/mL), samples were combined with a premixed cocktail of click reagents consisting of TEV tags (4μl of 5 mM stock, final concentration= 100 μM), TCEP (4 μl of fresh 50 mM stock in water, final concentration 1 mM), TBTA (12 μl of 1.7 mM stock in DMSO/t-butanol 1:4, final concentration = 100 μM), and CuSO4 (4 μl of 50 mM stock in water, final concentration = 1 mM). After 1h, the samples were then combined pairwise (400 μL total) and treated with 40 μL of 10%

SDS (1% SDS final) followed by 0.5 μL of benzonase (Novagen™ Benzonase™ Nuclease, Purity >90% MilliporeSigma™ 707464). Samples were left to incubate for 30 min at 37 °C. Following benzonase treatment, samples were subjected to the previously reported Single-Pot Solid-Phase-enhanced sample-preparation (SP3) protocol.[12,32]

Single-Pot Solid-Phase-enhanced sample-preparation (SP3).

40 μL Sera-Mag SpeedBeads Carboxyl Magnetic Beads, hydrophobic (GE Healthcare, 65152105050250) and 40 μL Sera-Mag SpeedBeads Carboxyl Magnetic Beads, hydrophilic (GE Healthcare, 45152105050250) were mixed and washed with water three times. The bead slurries were then transferred to the lysate, incubated for 10 min at RT with shaking (1000 rpm). 400uL of 200 proof EtOH was added to each sample and the mixtures were incubated for 5 min at RT with shaking (1000 rpm). The beads were then washed (2 × 1 mL 80% EtOH) with a magnetic rack. Proteins were eluted from SP3 beads with 200 μL of 0.5% SDS in PBS for 30 min at 37 °C with shaking (1000 rpm). 10 μL of 200 mM DTT (10 μM final concentration) was then added to each sample and the samples incubated at 65 °C for 15 min. Following reduction, 10 μL of 400 mM iodoacetamide (20 μM final concentration) was added to each sample and the samples incubated for 30 min at 37 °C with shaking at 300 rpm. Subsequently, 600 μL of 200 proof ethanol was added to each sample, and the samples were incubated for 5 min at ambient temperature with shaking (500 rpm). Beads were then washed three times with 80 % ethanol in water. Samples were then diluted with 150 μL 2M urea in PBS followed by the addition of reconstituted MS grade trypsin (2 μg, Promega, V5111). The samples were subjected to water bath sonication for 1 min and subsequently left to digest overnight (16 - 18hr) at 37°C and shaking at 200 rpm. The digested peptide solution and SP3 beads were then transferred into 15 mL falcon tubes. Peptides were then rebound to SP3 beads via the addition of 3.8 mL of 100% acetonitrile for a final percentage of

≥95% acetonitrile by volume and the peptides were subjected to shaking at 1000 rpm for 10 minutes at ambient temperature. Beads were collected using a magnet and solution was discarded. Samples were washed with 1 mL of 100% acetonitrile three times. Digested peptides were then eluted with 100 µL of 2% DMSO in water, shaking at 1000 rpm for 30 min at 37°C. Supernatant was collected in a 1.5 mL centrifuge tube on ice after separating SP3 beads with a magnetic rack. SP3 beads were resuspended with an additional 100 µL of 2% DMSO in water, shaking at 1000 rpm for 45 min at 37°C. Supernatant was collected after separating SP3 beads with a magnetic rack and combined with the previous elution volume (200 µL total).

Streptavidin enrichment of labeled proteins.

Pierce^{TM} Streptavidin Agarose resin (Thermo Scientific^{TM}, PI20353) (100 µL of resin) was first washed 3x in 10 mL of PBS by centrifugation at 1,800 x *g* for 3 min per wash. Solution was aspirated, making sure not to disturb spun down resin. After washing, resin was resuspended in 1 mL PBS/sample and re-distributed into 1.5 mL microcentrifuge tubes. The 200 µL peptide elution from previously prepared SP3 method was then added to the 1 mL of PBS/streptavidin resin. Samples were enriched by rotation for 2h at ambient temperature. After enrichment, the resin was collected by centrifugation at 1,400 x *g* for 5 min, and supernatant was aspirated and discarded. The resin was then subjected to washes 2x in 1 mL of PBS and 2x in 1 mL of water by centrifugation at 1,400 x *g* for 5 min per wash. After carefully aspirating and discarding the water, the resin/peptide slurry was then treated with a TEV protease following the "TEV digestion" protocol.

TEV digestion of labeled peptides.

Following streptavidin enrichment, samples were resuspended in 75 µL TEV buffer (50 mM Tris, pH 8, 0.5 mM EDTA, 1 mM DTT). To the resuspended samples, 1.5 µL TEV protease

(2 mg/mL or 70 μM; MacroLab, UC Berkeley) was added and the reactions were rotated for 7h at 30°C. The samples were then harvested by centrifugation (3,000 x *g*, 1 min) and the supernatant was collected. The collected peptides were then desalted using Pierce^{TM} C18 100 μL Tips (Thermo Scientific^{TM}, 87784) following the manufacturer's protocol. Briefly, 10 mL of the following four solutions were prepared; A) 100% acetonitrile, B) 50:50 acetonitrile:ultra pure water, C) Ultra pure water with 0.1% trifluoroacetic acid, and D) 60% acetonitrile with 0.1% trifluoroacetic acid in ultrapure water. Each C18 100 μL tip was first equilibrated with 100 μL with solution A for a total of two times followed by equilibration with solution B for a total of two times. The tips were then washed with 100 μL of solution C for a total of three times. Finally, 100 μL of samples were loaded into the tips and subsequently washed 2x with solution C. Samples were then eluted with 100 μL of solution D. Following desalting, each 100 μL sample was dried by speedvac, reconstituted in 20 μL of 5% acetonitrile and 1% formic acid in water, and analyzed by LC-MS/MS.

Liquid-chromatography tandem mass-spectrometry (LC-MS/MS) analysis.

The samples were analyzed by liquid chromatography tandem mass spectrometry using a Thermo Scientific™ Orbitrap Eclipse™ Tribrid™ mass spectrometer. Peptides were fractionated online using a 18 cm long, 100 μM inner diameter (ID) fused silica capillary packed in-house with bulk C18 reversed phase resin (particle size, 1.9 μm; pore size, 100 Å; Dr. Maisch GmbH). The 70-minute water-acetonitrile gradient was delivered using a Thermo Scientific™ EASY-nLC™ 1200 system at different flow rates (Buffer A: water with 3% DMSO and 0.1% formic acid and Buffer B: 80% acetonitrile with 3% DMSO and 0.1% formic acid). The detailed gradient includes 0 – 5 min from 3 % to 10 % at 300 nL/min, 5 – 64 min from 10 % to 50 % at 220 nL/min, and 64 – 70 min from 50 % to 95 % at 250 nL/min buffer B in buffer A. Data was collected with charge exclusion (1, 8, >8). Data was acquired using a Data-Dependent Acquisition (DDA) method consisting of a full MS1 scan (Resolution = 120,000) followed by sequential MS2 scans (Resolution = 15,000) to utilize the remainder of the 1 second cycle time. Precursor isolation window was set as 1.6 and normalized collision energy was set as 30%. The MS data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD056064.

Protein, peptide, and cysteine identification.

Raw data collected by LC-MS/MS were converted to mzML and searched with MSFragger (v3.3) and FragPipe (v19.0).[33,34,79,80] The proteomic workflow and its collection of tools was set as default and PTMprophet was enabled.[81] Precursor and fragment mass tolerance was set as 20 ppm. Missed cleavages were allowed up to 1. Peptide length was set 7 - 50 and peptide mass range was set 500 - 5000. For identification, cysteine residues were searched with differential modification C+. For cysteine reactivity quantification, MS1 labeling quant was enabled with Light set as C+521.3074 and Heavy set as C+527.3213. MS1 intensity ratio of heavy and light labeled cysteine peptides were reported with Ionquant (v1.8.9).[34] Calibrated and deisotoped spectrum files produced by FragPipe were retained and reused for this analysis. The MS search data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD056064. MS2 spectra data were searched using a reverse concatenated, non-redundant variant of the Human UniProtKB database (release-2020_01). Custom python scripts were implemented to compile labeled peptide datasets. Unique proteins, unique cysteines, and unique peptides were quantified for each dataset. Unique proteins were established based on UniProtKB protein IDs. Unique peptides were found based on sequences containing a modified cysteine residue. Unique cysteines were classified by an identifier consisting of a UniProtKB protein ID and the amino acid number of the modified cysteine

(UniProtKBID_C#); residue numbers were found by aligning the peptide sequence to the corresponding UniProtKB protein sequence. When there are multiple cysteines in one peptide, all the modified cysteine residue numbers will be reported as UniProtKBID_C#. Unique cysteines that were not found in at least two replicates and had an average isoTOP-ABPP ratio greater than three were discarded from further analysis.

3.4.2 - Computational Methods

Identification of High-reactive Cysteines in CysDB v1.5

Cysteines identified as high-reactive towards IAA from isoTOP-ABPP cysteine chemoproteomics experiments were downloaded from CysDB[30] (2401-release). CysteineIDs of "high-reactive" cysteines were cross-referenced with the list of cysteine identifiers in **Data S1**.

Active Site, Binding Site, Catalytic Activity, Disulfide Bond and Redox Potential Annotations of Detected Cysteines

Counts of how many proteins had UniProtKB annotations for active sites, binding sites, catalytic activity, disulfide bonds and redox potentials were calculated. Further parsing of UniProtKB active and binding site annotations were extracted to obtain specific residues and amino acid numbers. Positions of binding and active sites that were not cysteine residues were discarded. Exact amino acid positions of UniProtKB cysteine active and binding sites were cross-referenced with cysteine identifiers.

Cysteines Resolved in Crystal Structures

UniProtKB proteins from cysteines identified in **Data S1** were mapped to their corresponding Protein Data Bank[82] (PDB) identifiers. Peptides associated with reactive cysteines from **Data S1** were extracted for mapping to the corresponding PDB residue numbers. PDB structures were programmatically downloaded and parsed using the fetchPDB() and parsePDB() functions from the Protein Dynamics & Sequence Analysis[83] (ProDy) python package. To identify peptide sequences of reactive cysteines that were resolved in their associated PDB structures, fragments of each peptide (i - 3 : i : i + 3) were mapped to the residues in the associated structure. Lists of peptide fragments found in associated PDB structures were stored.

PDB Filtering

From the list of peptide fragments found in PDB structures, these PDBs were parsed again to obtain further information including, protein modifications, mutations, engineered mutations, experimental determination method and resolution (for X-ray crystallography and electron microscopy structures). PDB sequences were also inspected for "completeness." The number of disordered residues, designated by an "X", at the N- and C-termini were counted. If the sum of disordered residues at the N- and C-termini were equivalent to the total number of disordered residues within the structure, the PDB was classified as "complete." For these "complete" PDBs, a new PDB structure was written, removing any water molecules and crystallographic additives.

Each "complete" PDB was analyzed for potential heteroatoms, ions, and protein or nucleic acids within 6 Å of the reactive cysteine's resolved SG atom. The presence of alternative locations for side chains within 6 Å were noted, as well as the completeness of the side chains within 6 Å. These structures were then cross-referenced with a pre-compiled list of non-redundant protein chains from PISCES [37] (accessed November 2023). Representative chains were chosen based on either the highest resolution for X-ray crystallography or R-values for nuclear magnetic resonance spectroscopy (NMR) structures, where X-ray structures are given priority over other experimental determination methods.[37]

Criteria
No mutated or modified cysteine residue
Sequence complete (tolerance of 5 missing residues on N-, C-terminus, or both)
No heteroatoms (including ions, except Zn) nearby (at < 6 Å of SG atom)
No nucleic acids nearby (at < 6 Å of SG atom)
No altloc for sidechains (at < 6 Å of SG atom)
Nearby side chains complete (at < 6 Å of SG atom)

Functional Annotations of Detected Proteins

Custom Python scripts classified protein functions based on annotations in the UniProtKB/Swiss-Prot database (2301-release), Human Protein Atlas[84] (HPA) version 21.1 and the ScaPD[85] database. UniProtKB keywords were parsed to classify proteins into five broad functional categories: chaperones/transporter/channel/receptor, enzyme, nucleic acid and small molecule binding, scaffolding/modulator/adaptor, transcription factor/regulator and uncategorized. Transcription factors, channels and transporters were also found using protein class descriptions from the HPA. In addition, examples of experimentally validated scaffolding proteins were collected from the ScaPD database. For proteins in more than one category, annotations were prioritized based on the following: enzyme > chaperones/transporter/channel/receptor > scaffolding/modulator/adaptor > transcription factor/regulator > nucleic acid and small molecule binding.

Primary Sequence Logo

Peptides associated with high-reactive cysteines from **Data S1** were mapped to their corresponding canonical protein sequences using the UniProtKB reference FASTA (2301-release). Sequences were generated by fixing the high-reactive cysteine at position 0 and including the 10 amino acids upstream and downstream in the primary sequence, resulting in 21-amino-acid peptides. Starting with 823 high-reactive cysteines and 996 low-reactive cysteines, sequences were aligned to meet the input requirements of the pLogo software. The dataset was subsequently reduced to 765 high-reactive cysteines as the foreground and 805 low-reactive cysteines as the background. These sequences (n = 765) and their corresponding UniProtKB identifiers were compiled into a FASTA file for multiple sequence alignment (MSA) using CLUSTALW.[86] The MSA results were then used as input for pLogo.[35] The same process was applied to peptides associated with low-reactive cysteines (n = 805), which served as the background set. The sequence logo was generated using pLogo (http://plogo.uconn.edu).

Tertiary Structure Amino Acid Composition of Hyper-reactive Cysteines

For each cysteine resolved in a PDB structure, the coordinates of the reactive cysteine SG atom were identified. A list of all atoms within 5 Å, 7.5 Å, and 10 Å of the reactive cysteine SG atom was generated, excluding atoms from the reactive cysteine itself. These neighboring atoms were grouped by their residue names and amino acid numbers, and only unique residue identifiers (PDB_Chain_C#) were retained to avoid overcounting the same residue. If a residue had at least one atom within the specified radius, it was counted only once as a unique entity, ensuring a non-redundant representation of the local environment.

To determine the amino acid composition, we calculated the ratio of each of the 20 amino acids in both the high-reactive and low-reactive sets. For the high-reactive set, each amino acid was counted once per residue within the 7.5 Å radius and divided by the total number of unique residues in the set. This process was repeated for each of the 20 amino acids. The same procedure was applied to the low-reactive set. Finally, the Log2 ratio of each amino acid's distribution in the high-reactive set to its distribution in the low-reactive set was calculated. This allowed us to compare the differences in amino acid content between the two sets using 3D protein structures.

A 5 Å cutoff provided a larger number of significant proximal relations, while a 10 Å cutoff dramatically reduced the number of proximal connections. Therefore, we continued using a 7.5 Å cutoff to balance the need for proximal relations without overwhelming the data with noise.

Cysteines in Secondary Structure Motifs

The Dictionary of Secondary Structure-2[38,39] (DSSP-2) algorithm was used to classify the secondary structure motifs of each reactive cysteine. DSSP codes H ("alpha-helix"), G ("3-helix"), and I ("5-helix") were grouped as motifs pertaining to helices. DSSP codes B ("residue in isolated β-bridge") and E ("extended strand, participates in β ladder") were grouped as motifs pertaining to beta sheets. DSSP codes T ("hydrogen bonded turn") and S ("bend") were grouped as motifs pertaining to loops. Cysteines resolved in PDB structures that had DSSP codes in two or more of the three groups were categorized as "Conflict" and cysteines resolved in PDB structures that did not have a DSSP code were categorized as "NA."

Secondary Structure	DSSP Code	Category
Alpha helix (4-12)	H	Helix
Isolated beta-bridge residue	B	Beta Sheet
Strand	E	Beta Sheet
3-10 helix	G	Helix
Pi helix	I	Helix
Turn	T	Loop
Bend	S	Loop
None	-	NA

Cysteines in Zinc-Binding Motifs

Counts of how many proteins had UniProtKB annotations for binding sites were calculated. Further parsing of UniProtKB binding site annotations were extracted to obtain specific residues, amino acid numbers, and ligands. Positions of binding sites that did not have zinc as the associated ligand were discarded. Exact amino acid positions of UniProtKB zinc binding sites were cross-referenced with cysteine identifiers.

Predicted pKa values of Cysteines Resolved in PDB Structures

PROPKA[18] v3.1 was used to compute the predicted pKa value of each reactive cysteine resolved in an associated crystal structure. For each UniProtKBID_C# identifier, the median predicted pKa was computed.

Relative Solvent Accessibility of Cysteines Resolved in PDB Structures

DSSP-2 was used to compute the relative solvent accessibility (RSA), based on the Kabsch and Sander method, of each reactive cysteine resolved in an associated crystal structure. For each UniProtKBID_C# identifier, the median RSA was computed.

Pocket detection of high-reactive cysteines of Cysteines Resolved in PDB Structures

Fpocket[40] release 4.2 was used to detect ligand-binding pockets in each of the training set PDB structures. For each PDB, the resulting pocket folders were queried to extract all cysteine residues. A cysteine, based on a UniProtKBID_C# identifier, was classified as “in a predicted pocket” if it appeared in at least one predicted pocket.

Residue 1D Proximity Descriptors

Peptides associated with high-reactive cysteines from **Data S1** were mapped to their corresponding canonical protein sequences using the UniProtKB reference FASTA (2301-release). Sequences were generated by fixing the high-reactive cysteine at position 0. The distance (in number of amino acids) from the high-reactive cysteine at position 0 to the first occurrence of another cysteine, upstream in the primary sequence, was defined as the "1D_prox_left" descriptor. Similarly, the distance to the first occurrence of another cysteine downstream in the primary sequence was defined as the "1D_prox_right" descriptor. In both cases, the other cysteine had to be distinct from the high-reactive cysteine itself.

Secondary Structure Motif Descriptors

The DSSP-2 algorithm was used to classify the secondary structure motifs of each reactive cysteine. DSSP codes H ("alpha-helix"), G ("3-helix"), and I ("5-helix") were grouped as motifs pertaining to helices. DSSP codes B ("residue in isolated β-bridge") and E ("extended strand, participates in β ladder") were grouped as motifs pertaining to beta sheets. DSSP codes T ("hydrogen bonded turn") and S ("bend") were grouped as motifs pertaining to loops. Cysteines resolved in PDB structures that had DSSP codes in two or more of the three groups were categorized as "Conflict," and cysteines without a DSSP code were categorized as "NA."

The helix, beta sheet, and loop categories were one-hot encoded, converting the secondary structure information into a binary format with separate descriptors for helix, beta sheet, and loop.

Residue 3D Proximity Descriptors

For each cysteine resolved in a PDB structure, the coordinates of the reactive cysteine's SG atom were identified. A list of all atoms within 7.5 Å of the reactive cysteine's SG atom was generated, excluding atoms from the reactive cysteine itself. Atoms from non-cysteine residues were discarded. The Euclidean distance between the SG atom of each neighboring cysteine and the SG atom of the reactive cysteine was then calculated. The shortest distance was defined as the "3D_prox" descriptor.

Amino Acid Content Descriptors

For each cysteine resolved in a PDB structure, the coordinates of the reactive cysteine SG atom were identified. A list of all atoms within 7.5 Å of the reactive cysteine SG atom was generated, excluding atoms from the cysteine itself. These neighboring atoms were grouped by their residue names and amino acid numbers, and only unique residue identifiers (PDB_Chain_C#) were retained to avoid overcounting the same residue.

To determine the amino acid composition, the occurrences of each of the 20 amino acids were counted and divided by the total number of unique residue identifiers (PDB_Chain_C#) within 7.5 Å of the reactive cysteine SG atom.

Amino Acid Type Descriptors

For each cysteine resolved in a PDB structure, the coordinates of the reactive cysteine SG atom were identified. A list of all atoms within 7.5 Å of the reactive cysteine SG atom was generated, excluding atoms from the cysteine itself. These neighboring atoms were grouped by their residue names and amino acid numbers, and only unique residue identifiers (PDB_Chain_C#) were retained to avoid overcounting the same residue.

Amino acid type descriptors were assigned based on the residue and atom properties described by Cheng et al.[42] Each residue in the reactive cysteine neighborhood was classified into one of the following categories: polar (P), acidic (A), basic (B), aliphatic (ALI), and aromatic (ARO). The number of atoms in each category was counted and divided by the total number of residues within the 7.5 Å cutoff of the reactive cysteine SG atom, yielding the percentage of P (or other category) residues in relation to the total residues in the cysteine's neighborhood.

Atom Type Descriptors

For each cysteine resolved in a PDB structure, the coordinates of the reactive cysteine SG atom were identified. A list of all atoms within 7.5 Å of the reactive cysteine SG atom was generated, excluding atoms from the cysteine itself. Amino acid interaction descriptors were assigned based on the residue and atom properties described by the Graph-based Residue neighborhood Strategy to Predict binding sites (GRaSP)[43] method. Each atom in the reactive cysteine neighborhood was classified into one of the following categories: hydrophobic (HPB), donor (DON), acceptor (ACP), aromatic (ARO), positive (POS), negative (NEG), sulfide bridge (SSB), or "-". The number of atoms in each category was counted and divided by the total number of atoms within the 7.5 Å cutoff of the reactive cysteine SG atom, yielding the percentage of HPB (or other category) atoms in relation to the total atoms in the cysteine's neighborhood.

Residue	Atom	Property	Residue	Atom	Property	Residue	Atom	Property
ALA	N	DON	HIS	O	ACP	SER	CB	-
ALA	CA	-	HIS	CB	HPB	SER	OG	DON,ACP
ALA	C	-	HIS	CG	ARM	THR	N	DON
ALA	O	ACP	HIS	ND1	ARM,POS,DON,ACP	THR	CA	-
ALA	CB	HPB	HIS	CD2	ARM	THR	C	-
ARG	N	DON	HIS	CE1	ARM	THR	O	ACP
ARG	CA	-	HIS	NE2	ARM,POS,DON,ACP	THR	CB	-
ARG	C	-	ILE	N	DON	THR	OG1	DON,ACP
ARG	O	ACP	ILE	CA	-	THR	CG2	HPB
ARG	CB	HPB	ILE	C	-	TRP	N	DON
ARG	CG	HPB	ILE	O	ACP	TRP	CA	-
ARG	CD	-	ILE	CB	HPB	TRP	C	-
ARG	NE	POS,DON	ILE	CG1	HPB	TRP	O	ACP

ARG	CZ	POS	ILE	CG2	HPB	TRP	CB	HPB
ARG	NH1	POS,DON	ILE	CD1	HPB	TRP	CG	HPB,ARM
ASN	N	DON	LEU	N	DON	TRP	CD1	ARM
ASN	CA	-	LEU	CA	-	TRP	CD2	HPB,ARM
ASN	C	-	LEU	C	-	TRP	NE1	ARM,DON
ASN	O	ACP	LEU	O	ACP	TRP	CE2	ARM
ASN	CB	HPB	LEU	CB	HPB	TRP	CE3	HPB,ARM
ASN	CG	-	LEU	CG	HPB	TRP	CZ2	HPB,ARM
ASN	OD1	-	LEU	CD1	HPB	TRP	CZ3	HPB,ARM
ASN	ND2	DON	LEU	CD2	HPB	TRP	CH2	HPB,ARM
ASP	N	DON	LYS	N	DON	TYR	N	DON
ASP	CA	-	LYS	CA	-	TYR	CA	-
ASP	C	-	LYS	C	-	TYR	C	-
ASP	O	ACP	LYS	O	ACP	TYR	O	ACP

ASP	CB	HPB	LYS	CB	HPB	TYR	CB	HPB
ASP	CG	HPB	LYS	CG	HPB	TYR	CG	HPB,ARM
ASP	OD1	NEG,ACP	LYS	CD	HPB	TYR	CD1	HPB,ARM
ASP	OD2	NEG,ACP	LYS	CE	-	TYR	CD2	HPB,ARM
CYS	N	DON	LYS	NZ	-	TYR	CE1	HPB,ARM
CYS	CA	-	MET	N	DON	TYR	CE2	HPB,ARM
CYS	C	-	MET	CA	-	TYR	CZ	ARM
CYS	O	ACP	MET	C	-	TYR	OH	DON,ACP
CYS	CB	HPB	MET	O	ACP	VAL	N	DON
CYS	SG	DON,ACP,SSB	MET	CB	HPB	VAL	CA	-
GLN	N	DON	MET	CG	HPB	VAL	C	-
GLN	CA	-	MET	SD	ACP	VAL	O	ACP
GLN	C	-	MET	CE	HPB	VAL	CB	HPB
GLN	O	ACP	PHE	N	DON	VAL	CG1	HPB

GLN	CB	HPB	PHE	CA	-	VAL	CG2	HPB
GLN	CG	HPB	PHE	C	-			
GLN	CD	-	PHE	O	ACP			
GLN	OE1	ACP	PHE	CB	HPB			
GLN	NE2	DON	PHE	CG	HPB			
GLU	N	DON	PHE	CD1	HPB,ARM			
GLU	CA	-	PHE	CD2	HPB,ARM			
GLU	C	-	PHE	CE1	HPB,ARM			
GLU	O	ACP	PHE	CE2	HPB,ARM			
GLU	CB	HPB	PHE	CZ	HPB,ARM			
GLU	CG	HPB	PRO	N	DON			
GLU	CD	-	PRO	CA	-			
GLU	OE1	NEG,ACP	PRO	C	-			
GLU	OE2	NEG,ACP	PRO	O	ACP			

GLY	N	DON	PRO	CB	HPB
GLY	CA	-	PRO	CG	HPB
GLY	C	-	PRO	CD	-
GLY	O	ACP	SER	N	DON
HIS	N	DON	SER	CA	-
HIS	CA	-	SER	C	-
HIS	C	-	SER	O	ACP

Amino Acid Interaction Descriptors

For each cysteine resolved in a PDB structure, the coordinates of the reactive cysteine SG atom were identified. A list of all atoms within 7.5 Å of the reactive cysteine SG atom was generated, excluding atoms from the cysteine itself. Amino acid interaction descriptors were assigned based on the residue and atom properties described by the GRaSP[43] method. Each atom in the reactive cysteine neighborhood was classified into one of the following categories: hydrophobic (HPB), donor (DON), acceptor (ACP), aromatic (ARO), positive (POS), negative (NEG), sulfide bridge (SSB), or "-". Pairwise comparisons of all atoms in the reactive cysteine neighborhood were then made to identify potential aromatic stacking, hydrogen bonding, hydrophobic, repulsive, and salt bridge interactions. The number of interactions in each category was counted and divided by the total number of interactions within the 7.5 Å cutoff of the reactive cysteine SG atom, yielding the percentage of interactions in each category relative to the total number of interactions in the cysteine's neighborhood.

Interaction Type	Atom Types	Distance Min	Distance Max
Aromatic stacking	2 aromatic atoms	1.5	3.5
Hydrogen bond	1 acceptor atom and 1 donor atom	2	3
Hydrophobic	2 hydrophobic atoms	2	3.8
Repulsive	2 atoms with the same charge	2	6
Salt bridges	2 atoms with opposite charge	2	6

Amino Acid Steric Interaction Descriptor

For each cysteine resolved in a PDB structure, the coordinates of the reactive cysteine SG atom were identified. A list of all atoms within 7.5 Å of the reactive cysteine SG atom was generated, excluding atoms from the cysteine itself. A steric interaction was classified based on if the distance between the SG atom of the reactive cysteine and whether an atom in a neighboring residue was less than the sum of their Van Der Waals radii.[45] The steric_P_7.5 descriptor is the total number of steric interactions between the SG atom of the reactive cysteine and any other atom in a neighboring residue, divided by the total number of unique PDB_Chain_C# identifiers.

Atom	**Van Der Waals Radii**
H	1.2
C	1.7
N	1.55
O	1.52

Hydrophobicity Kyte-Doolittle Descriptor

For each cysteine resolved in a PDB structure, the coordinates of the reactive cysteine SG atom were identified. A list of all atoms within 7.5 Å of the reactive cysteine SG atom was generated, excluding atoms from the cysteine itself. These neighboring atoms were grouped by their residue names and amino acid numbers, and only unique residue identifiers (PDB_Chain_C#) were kept to avoid overcounting the same residue. The sum of hydrophobicity values for each neighboring residue according to the Kyte-Doolittle[44] scale, divided by the total number of neighboring residues within the distance cutoff, was computed to create the “hydrophobicity_kd” descriptor.

Amino Acid	**Acidity**	**Hydrophobicity**
ALA	ALI	1.8
ARG	B	-4.5
ASN	P	-3.5
ASP	A	-3.5
CYS	P	2.5
GLN	P	-3.5
GLU	A	-3.5
GLY	ALI	-0.4
HIS	B	-3.2
ILE	ALI	4.5
LEU	ALI	3.8
LYS	B	-3.9
MET	ALI	1.9
PHE	ARO	2.8
PRO	P	-1.6
SER	P	-0.8
THR	P	-0.7
TRP	ARO	-0.9

TYR	ARO	-1.3
VAL	ALI	4.2

Predicted Pocket Descriptor

Fpocket[40] release 4.2 was used to detect ligand-binding pockets in each of the training set PDB structures. For each PDB, the resulting pocket folders were queried to extract all cysteine residues. A cysteine, identified by its PDB_Chain_C# identifier, was classified as "in a predicted pocket" (1) if it appeared in at least one predicted pocket, or "not in a predicted pocket" (0).

Predicted pKa Descriptor

PROPKA v3.1 was used to compute the predicted pKa value of each reactive cysteine in its associated crystal structure. The "backbone hydrogen bond" energies from the pKa output files were summed for each unique PDB_Chain_C# identifier to create the "pKa_hb_bb" descriptor. The "sidechain hydrogen bond" and "backbone hydrogen bond" energies from the pKa output files were summed for each unique PDB_Chain_C# identifier to create the "pKa_sd" descriptor. Then, the values of the "pKa_hb_bb" descriptor and "pKa_hb_sd" descriptor were summed to create the "pKa_hb" descriptor.

Finally, the sum of the Coulombic interaction energies and the "pKa_hb" descriptor was used to create the "pKa_hb_elec" descriptor. "Buriedness" and "predicted_desolvation" values were extracted directly from the raw PROPKA output files.

B-factor and Disorder Descriptors

For each cysteine resolved in a PDB structure, the coordinates of the reactive cysteine SG atom were identified. The associated b-factor of the SG atom was obtained using the get_bfactor() function of BioPython's[41] Atom class. The disorder of the cysteine residue was obtained using the is_disordered() function of BioPython's Residue class.

Cysteines in Disulfides Descriptors

For each cysteine resolved in a PDB structure, the coordinates of the reactive cysteine's SG atom were identified. A list of all atoms within 7.5 Å of the reactive cysteine's SG atom was generated, excluding atoms from the reactive cysteine itself. Atoms from non-cysteine residues were discarded. The Euclidean distance between the SG atom of each neighboring cysteine and the SG atom of the reactive cysteine was then calculated. A cysteine, identified by its PDB_Chain_C# identifier, was classified as potentially participating in a "disulfide_bridge" (1) if any of the Euclidean distances were less than or equal to 3 Å, or "not in a disulfide_bridge" (0).

Hydrogen Bond Descriptors

For each cysteine resolved in a PDB structure, the coordinates of the reactive cysteine SG atom were identified. A list of all atoms within 7.5 Å of the reactive cysteine SG atom was generated, excluding atoms from the cysteine itself. Hydrogen bond descriptors were assigned based on the residue and atom properties described in the table below. Each atom in the reactive cysteine neighborhood that was considered a donor or acceptor, either from the backbone or side chain, was classified into one of the following categories: hydrogen_bond_donor_backbone, hydrogen_bond_donor_sidechain, hydrogen_bond_acceptor_backbone, or hydrogen_bond_acceptor_sidechain.

The number of atoms in each category within a 5 Å distance was computed and divided by the total number of atoms within the 5 Å neighborhood of the reactive cysteine. Additionally, the number of atoms in each category within the distance range of 5 Å to 7.5 Å was calculated and divided by the total number of atoms within the 7.5 Å neighborhood of the reactive cysteine. The hydrogen bond descriptors were calculated for both a 0 - 5 Å cutoff and a 5 - 7.5 Å cutoff to account for both close-range and distal hydrogen bond interactions.

Hydrogen Bond Donor Atoms

Residue	Atom	Residue	Atom
ARG	NH, NE	THR	OG1
ASN	ND2	TYR	OH
GLN	NE2	SER	OG
HIS	NE2, ND1	CYS	SG
LYS	NZ		
TRP	NE1		
BACKBONE (NOT PRO)	N		

Hydrogen Bond Acceptor Atoms

Residue	Atom	Residue	Atom
ASP	OD	HIS	NE2, ND1
ASN	OD1	CYS	SG
GLN	OE1		
GLU	OE		

SER	OG
THR	OG1
TYR	OH
BACKBONE	O

Physicochemical Descriptors

For each PDB in the training and test sets, energetic contributions of various physicochemical properties of each reactive cysteine resolved in an associated crystal structure were calculated using Rosetta.[49] Energy terms collected were reported on a “per_reside” basis using default talaris2013 weights for each term. Each of the following terms were computed for each residue in the structure based on the residue type and x, y, z coordinate positions: “fa_atr”, “fa_rep”, “fa_sol”, “fa_intra_rep”, “fa_intra_sol_xover4”, “lk_ball_wtd”, “fa_elec”, “pro_close”, “hbond_sr_bb”, “hbond_lr_bb”, “hbond_bb_sc”, “dslf_fa13”, “omega”, “fa_dun”, “p_aa_pp”, and “rama_prepro”.

Energy Term	**Description**
fa_atr	Lennard-Jones attractive forces between two atoms of different residues
fa_rep	Lennard-Jones repulsive forces between two atoms of different residues
fa_sol	Gaussian solvent-exclusion model for solvation free energy between atoms in different residues
fa_intra_rep	Lennard-Jones repulsive forces between two atoms in the same residue
fa_intra_sol_xover4	Gaussian solvent-exclusion model for solvation free energy between atoms in the same residue
lk_ball_wtd	Asymmetric solvation of ions
fa_elec	Coulombic electrostatic potential between two non-bonded, charged atoms
pro_close	Energy associated with proline ring conformational stain and impact on psi angle of preceding residue

hbond_sr_bb	Short-range backbone-backbone hydrogen bond energy
hbond_lr_bb	Long-range backbone-backbone hydrogen bond energy
hbond_bb_sc	Backbone-side-chain hydrogen bond energy
dslf_fa13	Disulfide bridge energy
omega	Backbone omega dihedral
fa_dun	Internal energy of sidechain rotamer
p_aa_pp	Probability of amino acid using the Dunbrack library and backbone torsion phi and psi angles
rama_prepro	Likelihood of backbone phi and psi angles for amino acid

Environment and Tools

All data preprocessing and machine learning tasks using Python 3.10 and the Scikit-learn library (v1.5.2).[87] Additional libraries such as pandas (v2.2.3)[88] and numpy (v2.1.3)[21] were used for data handling, and matplotlib (v3.9.2)[89] and seaborn (v0.13.2)[90] were used for visualization. Feature interpretation was performed using the SHAP library (v1.0.3),[91] which provided insights into the contributions of individual descriptors to the Random Forest model's predictions.

Data Preprocessing and Binary Classification

A supervised machine learning approach was used to classify cysteines as low-reactive or high-reactive toward Iodoacetamide Alkyne (IAA). Binary classification labels were assigned as follows: 0 for low-reactive and 1 for high-reactive cysteines. These labels were derived from experimental data (**Data S1**), with 297 low-reactive cysteines and 306 high-reactive cysteines included in the dataset. The balanced distribution ensured that class imbalance was not a concern. Descriptors used as features were normalized to ensure consistency. For amino acid composition features, normalization was achieved by dividing the number of amino acids of a given type within a specified distance cutoff by the total number of amino acids within the same cutoff, preserving relative proportions without scaling to a fixed range.

Algorithm Comparison

Multiple supervised learning algorithms were evaluated to identify the most effective model for the classification task. The algorithms tested included Random Forest (RF), K-Nearest Neighbors (KNN), Classification and Regression Trees (CART), Linear Discriminant Analysis (LDA), and Support Vector Machines (SVM). These algorithms were selected to provide diverse modeling approaches, including ensemble-based methods, distance-based classification, and linear separation techniques. Each algorithm was trained on the training dataset using the descriptors as input features and the binary classification labels as targets. Performance was assessed using accuracy and the area under the receiver operating characteristic curve (ROC AUC).

Initial comparisons revealed that while each algorithm exhibited strengths, further optimization was required to balance the true positive rate (TPR) and false positive rate (FPR). The Random Forest classifier was selected for further development due to its ability to handle complex, nonlinear relationships and its robustness to overfitting.

Cross-Validation and Test Set Sampling

The dataset was divided into independent training and test sets to ensure unbiased evaluation of the model. Training and test sets were curated from distinct experimental sources, ensuring no overlap between cysteines in the two datasets. All cross-validation procedures were performed exclusively within the training dataset.

Five-fold cross-validation was implemented using Scikit-learn's KFold function to evaluate model performance during training. The training data was split into five non-overlapping subsets, with four subsets used for training and the remaining subset used for validation in each iteration. To ensure reproducibility, the shuffle parameter was set to True, and a fixed random seed (random_state = 42) was applied. The Random Forest classifier was trained on the training subsets, and validation accuracy was computed for the held-out subset during each fold. The cross-validation accuracy averaged 0.81, indicating strong generalization performance on the training dataset.

To further validate model stability, a repeated test set sampling approach was applied. The Random Forest classifier, trained on the full training dataset, was evaluated on the independent test set over five iterations. During each iteration, predictions for the test set were compared to the true binary labels, and metrics such as accuracy, precision, and recall were computed. This approach yielded consistent test set performance, with an average accuracy of 0.66.

Hyperparameter Optimization

Hyperparameter tuning was conducted using a grid search with five-fold cross-validation, implemented through Scikit-learn's GridSearchCV function. The following hyperparameters were evaluated: the number of estimators (n_estimators) with values of 5, 100, 150, and 200; the maximum depth of the trees (max_depth) with values of None (allowing trees to grow until leaves were pure or contained fewer than the minimum required samples to split), 2, 4, 6, 8, and 10; the number of features considered at each split (max_features) with options of None (all features), "sqrt" (square root of the total number of features), and "log2" (logarithm base 2 of the number of features); and the splitting criterion (criterion) with values of "gini" and "entropy." Each combination of these parameters was evaluated on the training set $(\chi_{train}, y_{train})$ using five-fold cross-validation, ensuring robust performance testing across multiple data partitions. The grid search identified the optimal configuration as n_estimators = 150, max_depth = 6, max_features = "log2," and criterion = "gini." The final model, trained with this configuration, achieved a test set accuracy of 64%, precision of 60%, recall of 73%, and an area under the receiver operating characteristic curve (AUC) score of 64%.

Feature Selection and Model Refinement

Recursive feature elimination with cross-validation (RFECV) was used to refine the feature set by systematically removing descriptors with low predictive importance. The procedure was implemented using Scikit-learn's RFECV function with a Random Forest classifier as the estimator. Five-fold cross-validation and an accuracy scoring metric were applied. At each iteration, the least important features were removed based on their contribution to model performance, and the model was retrained on the reduced feature set. A minimum of two features was specified (min_features_to_select = 2) to prevent over-reduction. The stopping criterion for RFECV was based on identifying the number of features that achieved the highest average cross-validation accuracy across folds. The process identified the optimal number of features, which were subsequently used to train the final Random Forest model. The Random Forest classifier was initialized with a fixed random seed (random_state = 42) to ensure reproducibility. RFECV was applied to the training set $(\chi_{train}, y_{train})$, and the optimal number of features was determined based on cross-validated accuracy scores. The final feature set was then used to train the optimized Random Forest model.

Ablation Process

After the initial recursive feature elimination (RFE), descriptors were grouped into categories, as detailed in **Data S3**. Ablation studies were conducted to systematically assess the impact of these descriptor categories on model performance. The grouped descriptor categories included amino acid composition (AAC), hydrogen bond statistics, relative solvent accessibility (RSA), and others. For each ablation step, a single descriptor category was removed from the feature set, and the remaining features were used to train a Random Forest classifier.

The performance of each reduced model was evaluated using five-fold cross-validation on the training set, focusing on metrics such as true positive rate (TPR) and false positive rate (FPR). These metrics were selected to quantify the model's ability to correctly classify high-reactive cysteines (TPR) while minimizing incorrect classifications of low-reactive cysteines (FPR). By comparing the TPR and FPR across models trained with different combinations of descriptor categories, the contributions of each category were quantified.

The combination of categories that yielded the highest TPR and lowest FPR—amino acid composition (AAC), hydrogen bond statistics, and RSA—was selected for further refinement. This optimal combination of descriptor categories was then subjected to another round of hyperparameter optimization and recursive feature elimination to determine the final feature set. During this second RFE step, all features were retained, as none were eliminated based on their importance scores. The resulting optimized model incorporated 29 features, which were then used to train the final Random Forest classifier.

3.5 - Supporting Information

3.5.1 - Supplementary Figures

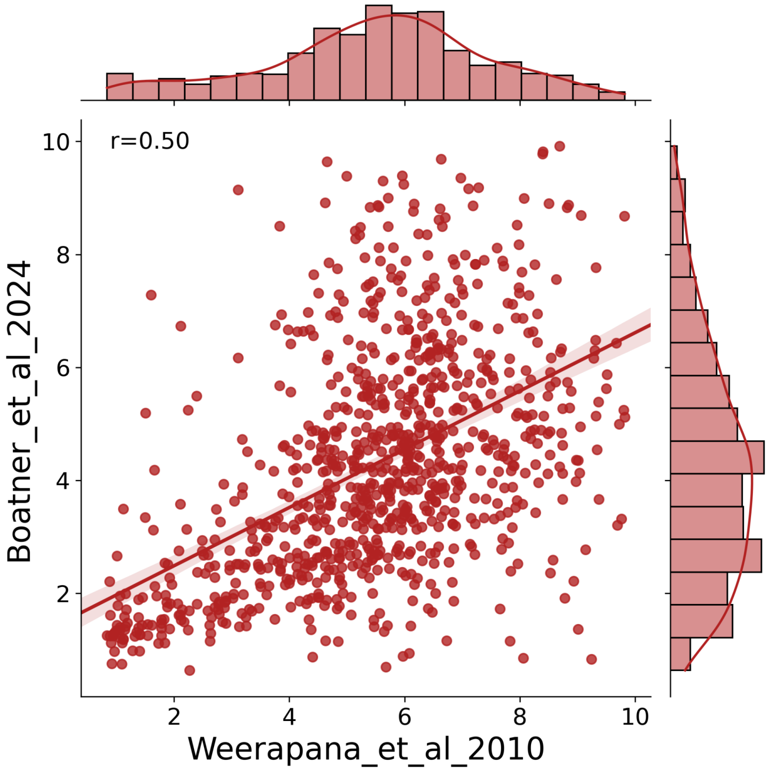


Figure S1. Total number of cysteine shared between Weerapana et al.[6] and this work, n = 915. Pearson correlation coefficient (*R*) = 0.5 and p-value < 0.001.

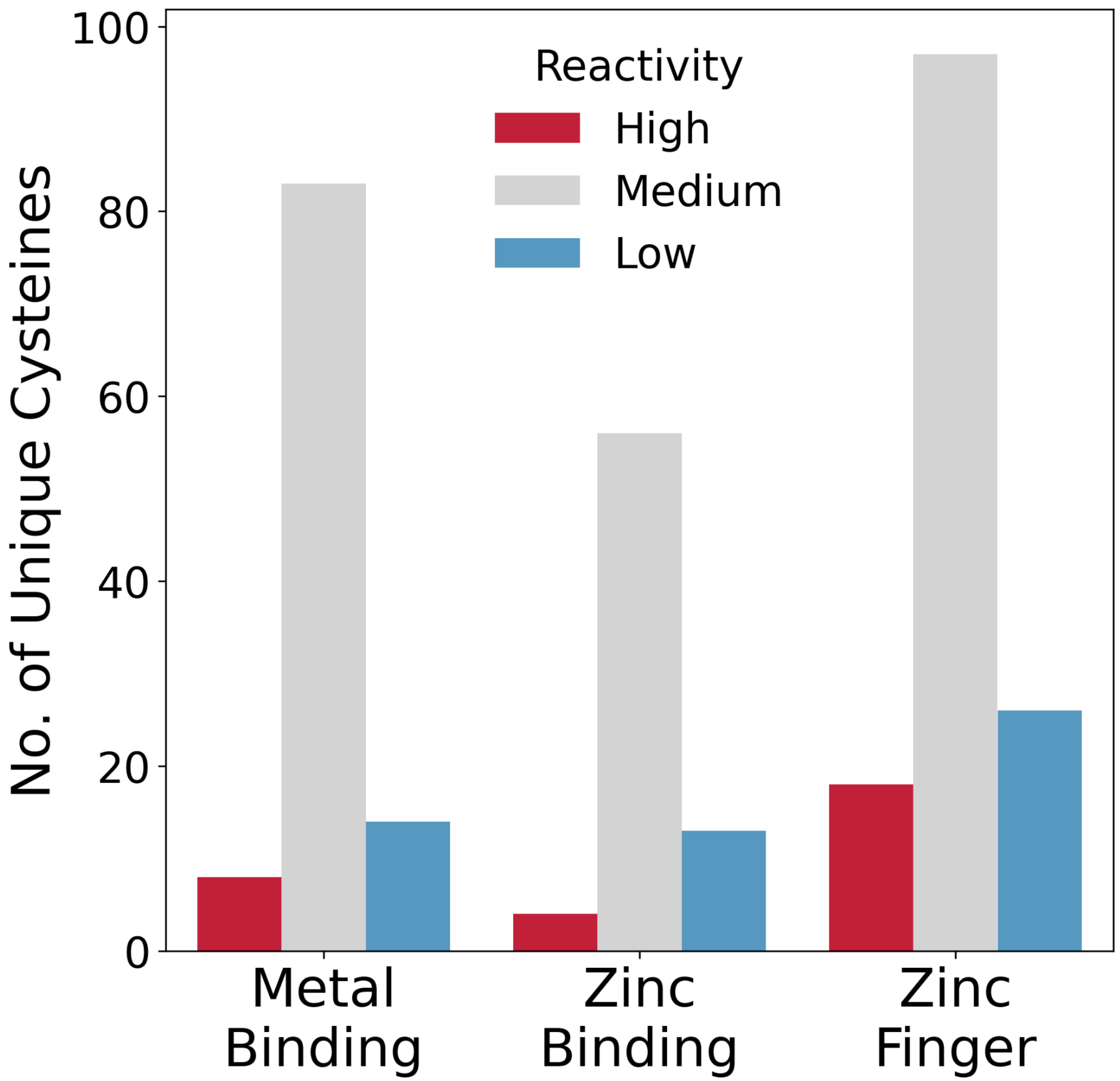


Figure S2. Number of IAA-reactive cysteines annotated by UniProtKB as metal binding sites (not including zinc), zinc binding sites, or within zinc finger regions. Further parsing of UniProtKB metal binding site annotations were extracted to obtain specific residues and amino acid numbers. Positions of metal binding site annotations that were not cysteine residues were discarded. Exact amino acid positions of UniProtKB cysteine metal binding sites were cross-referenced with IAA-reactive cysteine identifiers.

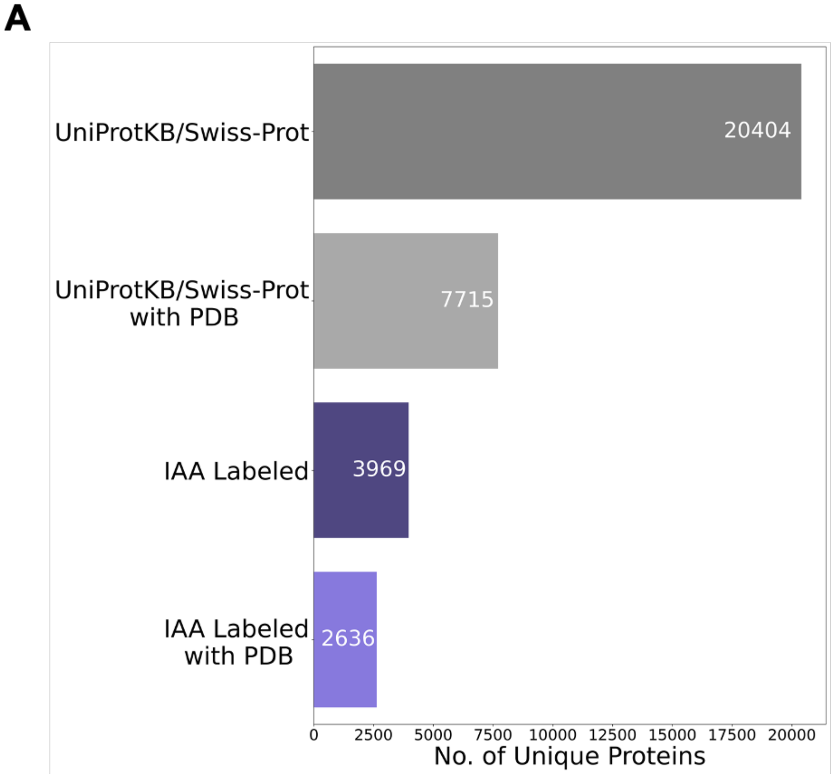


Figure S3. (A) Number of proteins in the entire human proteome and labeled by IAA vs number of proteins in the entire human proteome and labeled by IAA with associated PDB structures.

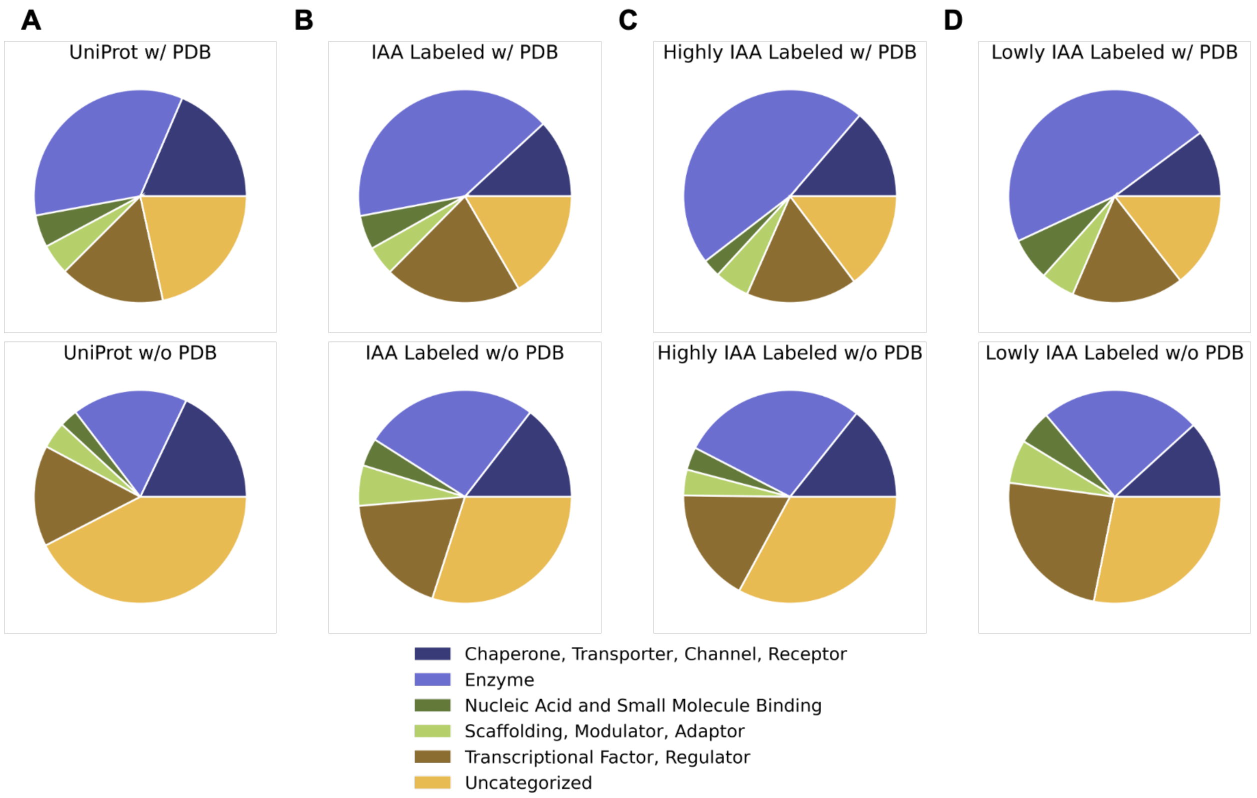


Figure S4. Functional classification of proteins in UniProtKB, experimentally identified, containing high-reactive cysteines, and containing low-reactive cysteines. Proteins associated with PDBs are found in the top row, while proteins without associated PDBs are found in the bottom row.

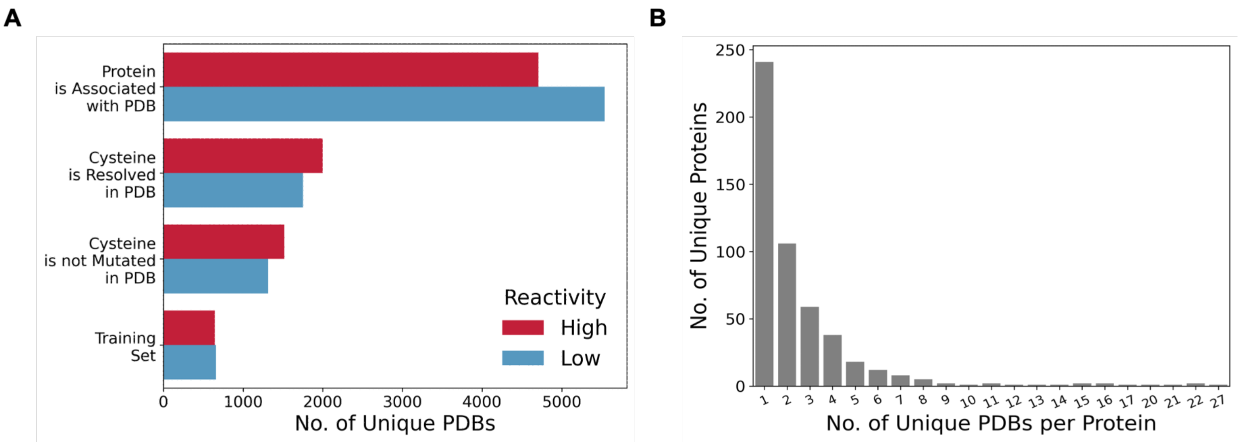


Figure S5. (A) Number of PDB structures associated with a UniProtKB protein (n = 505 proteins experimentally identified and have an associated PDB structure in the final training set). Average number of PDBs per UniProtKB protein is 2.6, with a standard deviation of 3.1. (B) Bar graph showing the number of unique PDB structures associated with proteins containing high- or low-reactive cysteines in the training set after a series of filtering steps.

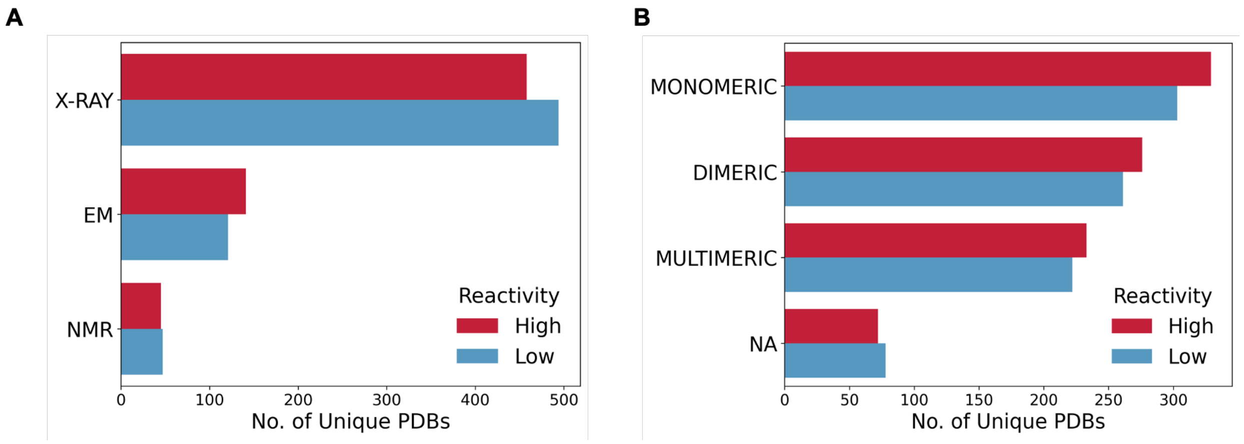


Figure S6. (A) Bar graph showing the number of unique PDB structures associated with proteins containing high- or low-reactive cysteines in the training set, categorized by protein structure determination methods. (B) Number of training set PDBs that are classified as monomeric, dimeric, multimeric or NA structures. A structure was considered mono-, di-, multi- or NA based on the reported biological unit annotated within the associated structure. Total number of monomeric biological units is 204, dimeric is 177, multimeric is 199, and NA is 72 for PDBs associated with proteins containing high-reactive cysteines. Total number of monomeric biological units is 218, dimeric is 185, multimeric is 183, and NA is 78 for PDBs associated with proteins containing low-reactive cysteines.

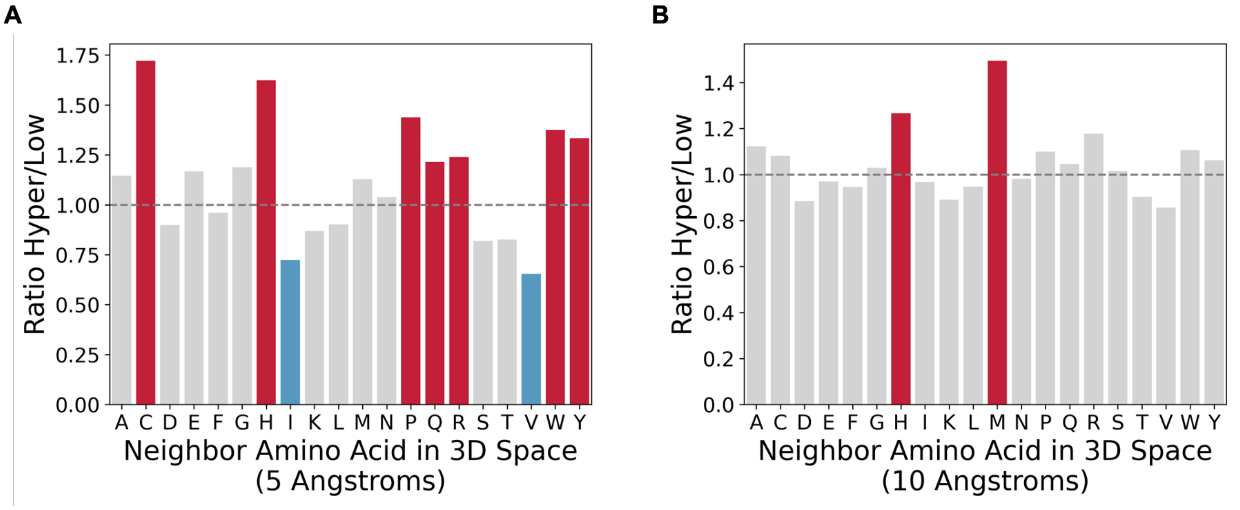


Figure S7. (A) Ratio of neighboring amino acids within 5 Å from the SG atom of a high- or low-reactive cysteine. (B) Ratio of neighboring amino acids within 10 Å from the SG atom of a high- or low-reactive cysteine. Each unique neighboring residue was counted only once. For each subset of PDB structures containing either high- or low-reactive cysteines, the number of each amino acid within the specified distance cutoff was computed. The relative frequencies of each amino acid in the high- or low-reactive neighborhoods were then divided by the total number of residues found in the respective neighborhoods across all associated PDBs. Finally, the ratio of high/low was calculated by dividing the relative frequency of each amino acid in the high-reactive neighborhoods by the relative frequency of each amino acid in the low-reactive neighborhoods.

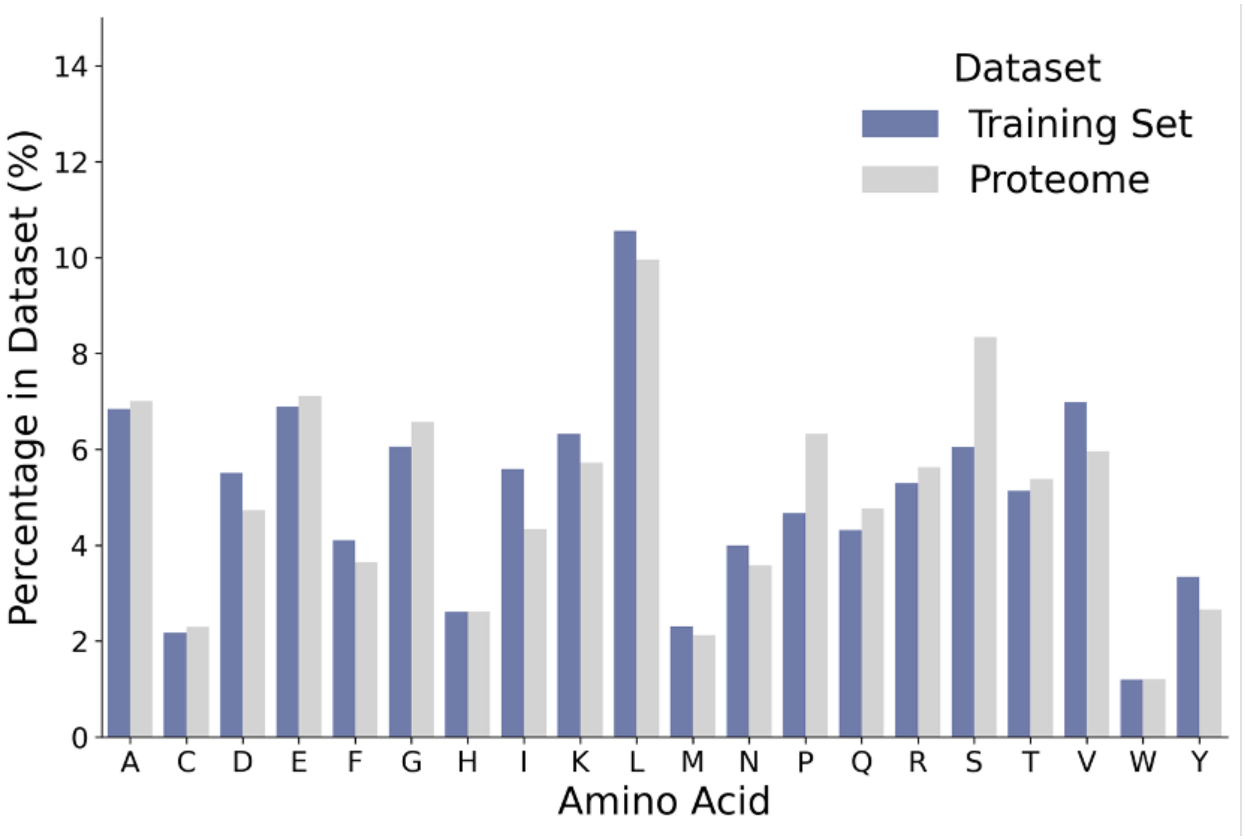


Figure S8. Distribution of amino acids in the human UniProtKB proteome and training set. Canonical protein sequences from the Human UniProtKB Proteome[92] (2301_release) were used to calculate the percentage (%) or abundance of each amino acid within the reference proteome. To calculate the amino acid percentage (%) or abundance using PDB structures in the training set, custom scripts collected the sequence of amino acids resolved in each PDB structure. The number of each amino acid was then divided by the total number of residues found within all training set PDB structures (n = 901,921).

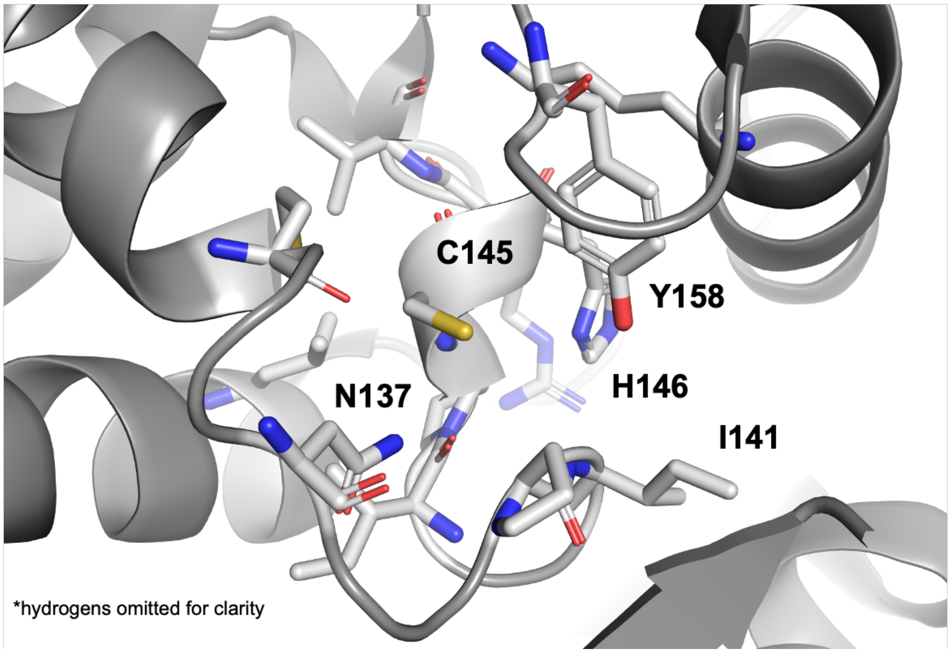
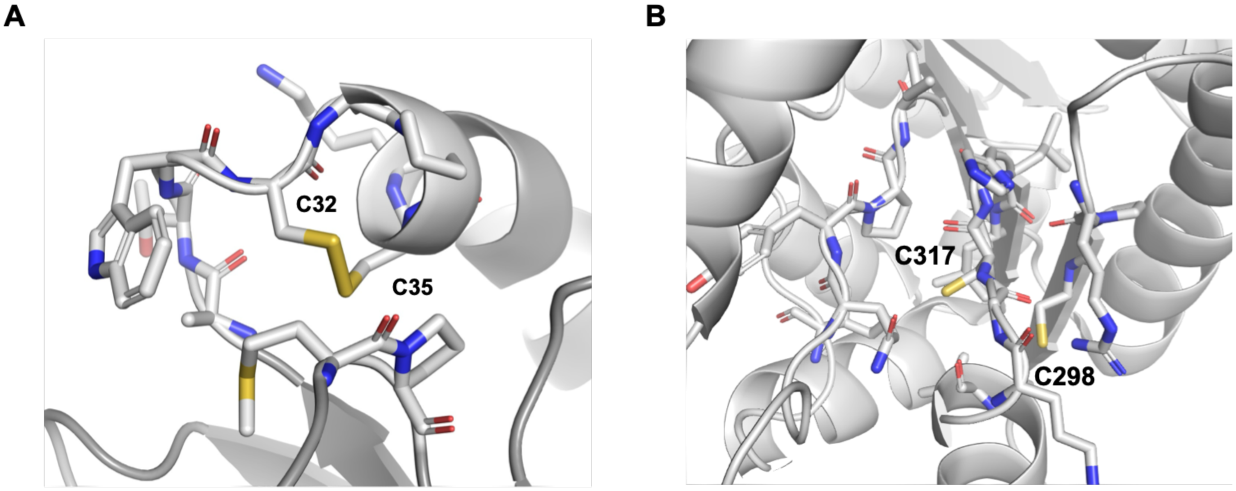


Figure S9. Close up view of O6-methylguanine-DNA methyltransferase, MGMT, C145 microenvironment (PDB: 1EH6). Experimental pKa was measured at 5.3.[50] Experimental isoTOP-ABPP ($R_{10:1}$) was measured at 0.87. Backbone hydrogen bond donors include I141 and H146, while side chain hydrogen bond donors include ASN137 and Y158.



*hydrogens omitted for clarity

*hydrogens omitted for clarity

Figure S10. (A) Close up view of Thioredoxin, TXN, C32 and C35 (PDB: 1ERU). Experimental evidence of disulfide bond between C32 and C35[52,55] was also observed in an associated crystal structure 1ERU. Median predicted pKa, using PROPKA[18] version 3.1, for C32 was calculated to be 10.92 and C35 was calculated to be 14. (B) Close up view of high-reactive cysteine C317 in ATP-dependent RNA helicase, DDX3X (PDB: 2I4I). Median predicted pKa, using PROPKA, for C317 was calculated to be 14. A cysteine was detected by custom scripts to potentially form a disulfide bond when another cysteine sulfur atom was within 3 Å of a reactive cysteine SG atom. Sulfur atom of C317 is within 5.1 Å from the sulfur atom of C298.

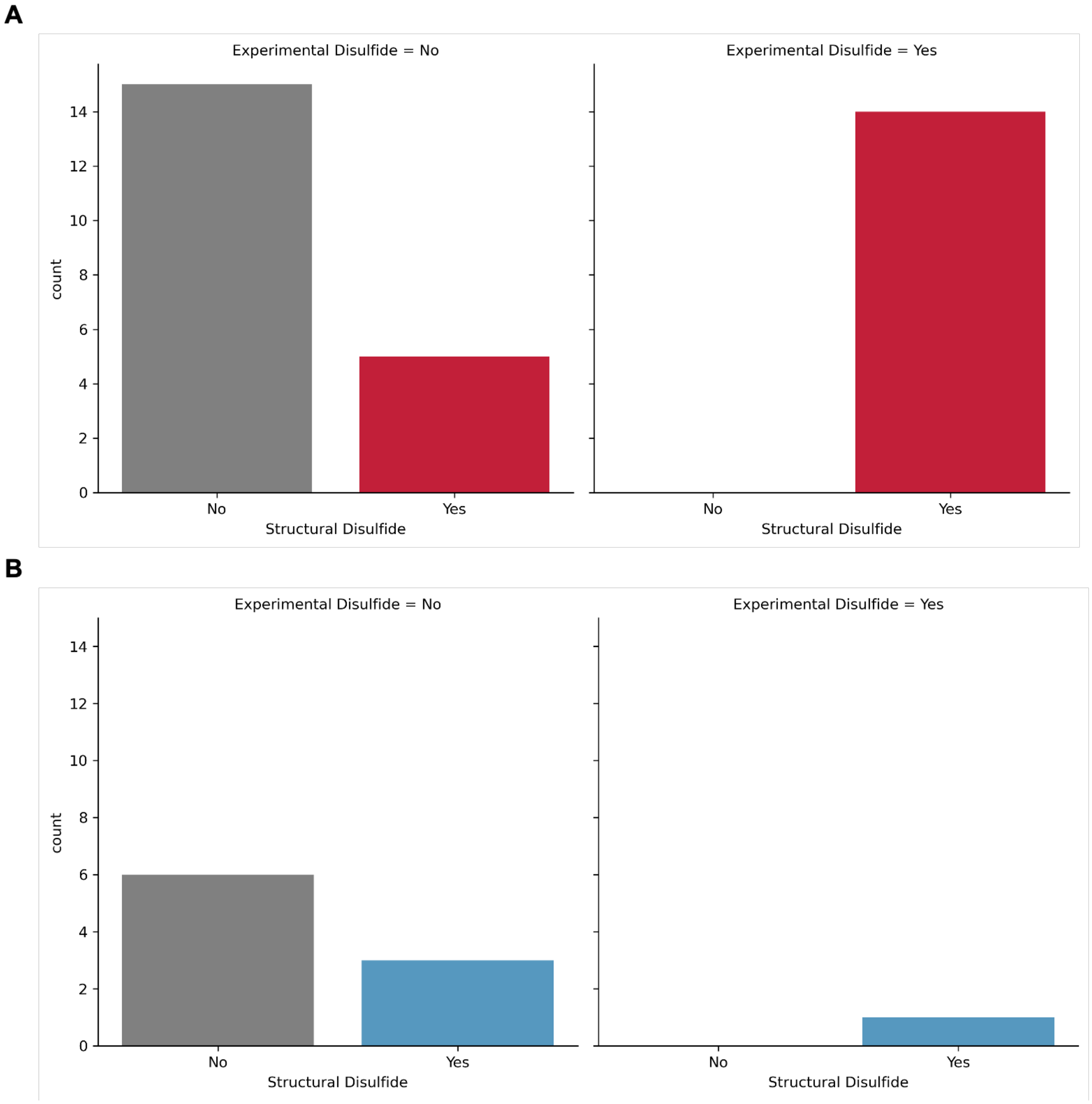


Figure S11. (A) Bar graph showing the number of high-reactive cysteines with experimental evidence of being involved in a disulfide bond using UniProtKB annotations. Additionally, a cysteine was considered a structural disulfide if it was found to be in a disulfide bond in an associated structure or when the sulfur atom of the high-reactive cysteine was within 3 Å of another cysteine SG atom. (B) Bar graph showing the number of low-reactive cysteines with experimental evidence of being involved in a disulfide bond using UniProtKB annotations. Additionally, a cysteine was considered a structural disulfide if it was found to be in a disulfide bond in an associated structure or when the sulfur atom of the low-reactive cysteine was within 3 Å of another cysteine SG atom. See **Data S1** and **Data S3**.

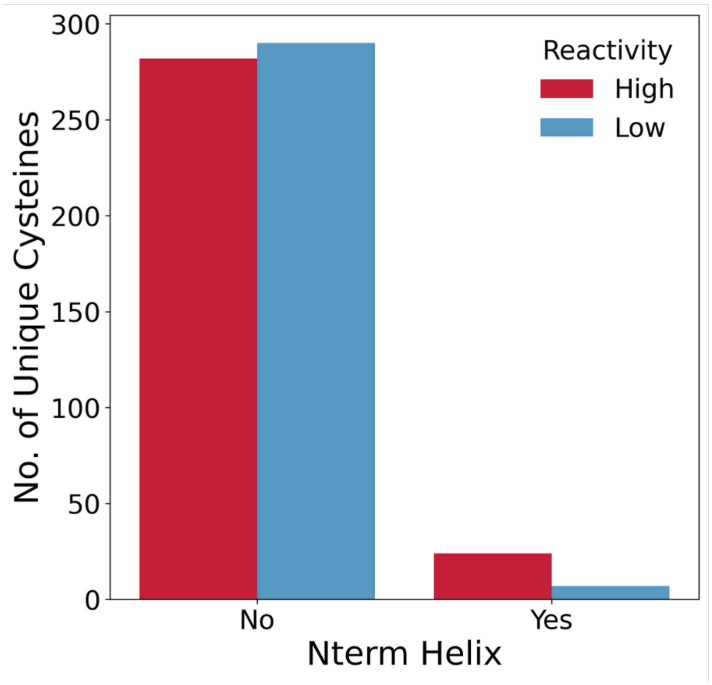


Figure S12. Number of IAA-reactive cysteines proximal to an N-terminal helix. A cysteine was defined as being near the N-terminus of a helix if the nitrogen atoms of the two downstream residues (i+2 and i+3)[16] were part of a helix and within 5 Å, even if the cysteine itself was not located within the helix.

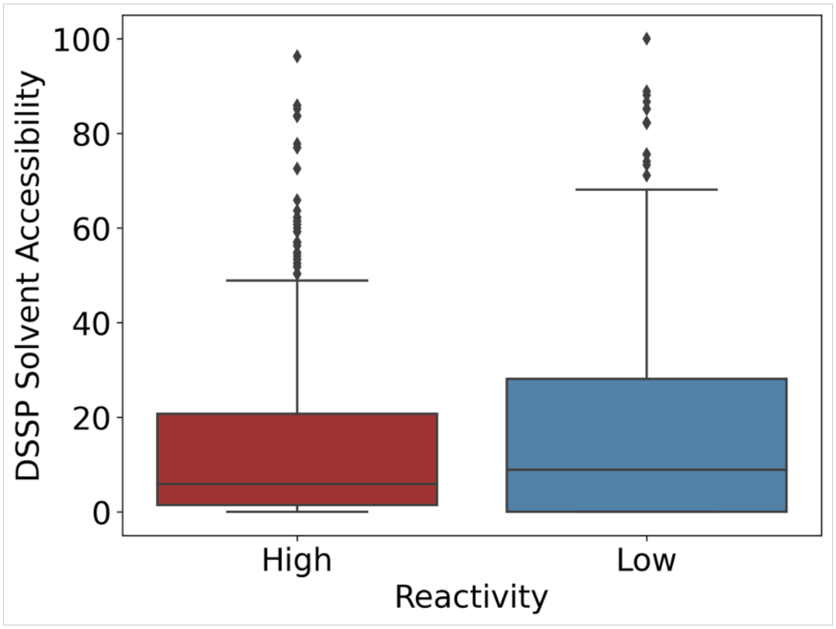


Figure S13. There is no relationship between computationally determined relative solvent accessibility (RSA) (DSSP-2) using predicted AlphaFold 2[57] structures and quantitative measurements of cysteine reactivity (Mann-Whitney U Test statistic: 25,515,605, p = 0.4859).

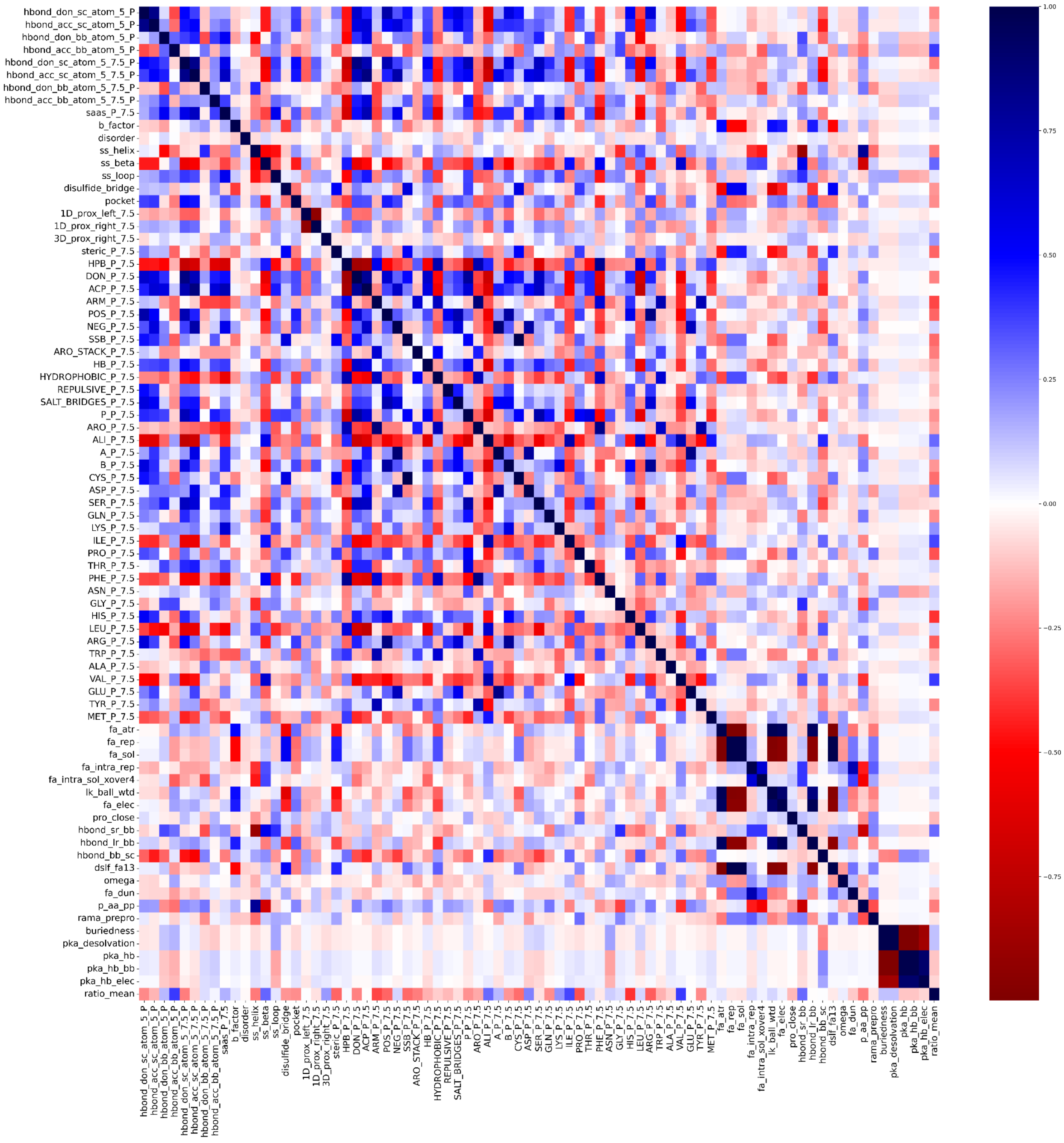


Figure S14. Heatmap of Pearson Correlation Coefficients (PCC) between measures of cysteine reactivity (**Data S1**) and descriptors obtained from 3D structures. See **Supplementary Computational Methods** and **Data S3**.

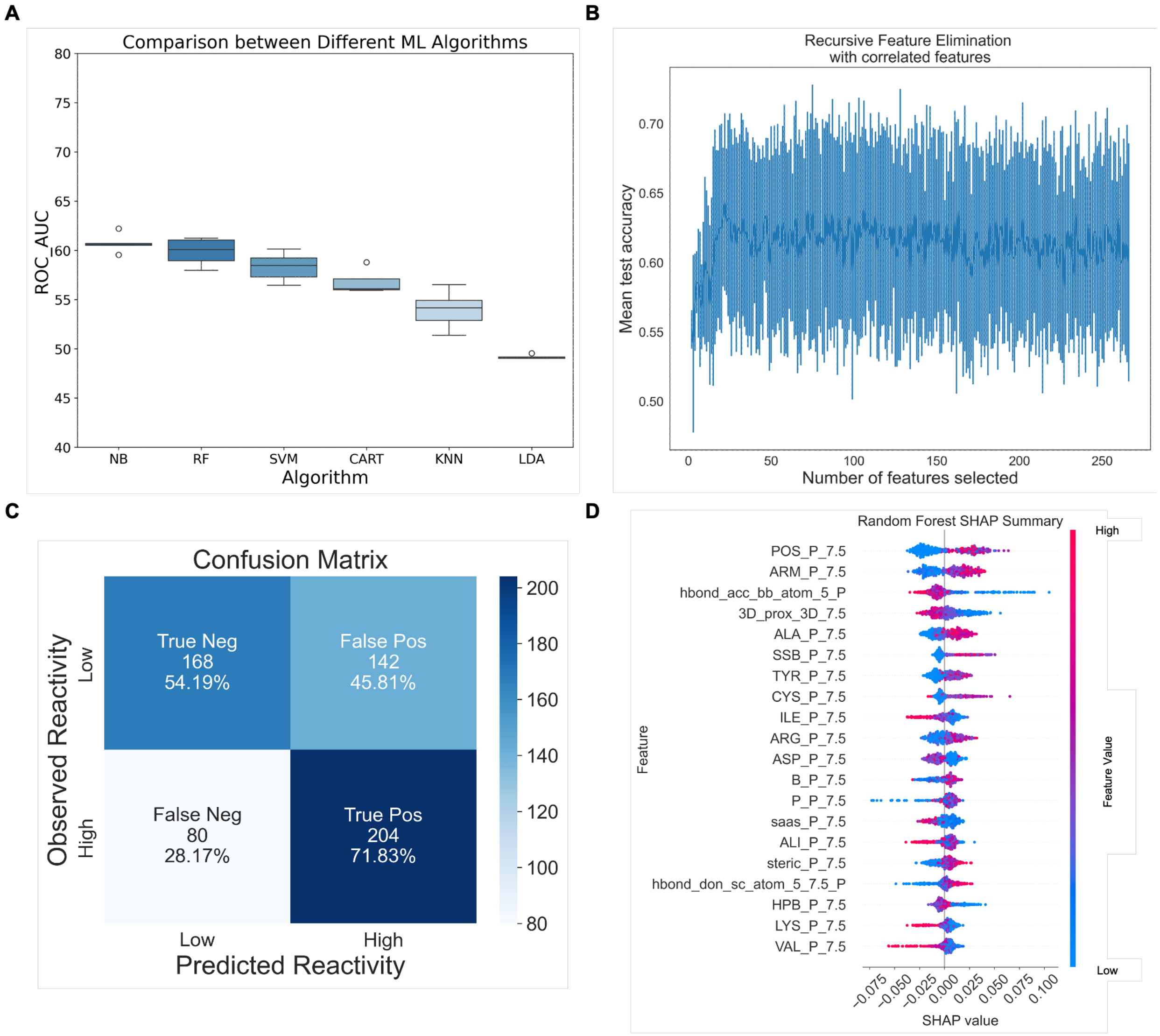


Figure S15. (A) Algorithm comparison of training data after training on isoTOP-ABPP 2024 experimental dataset and testing on aggregated dataset from the literature.[26,63–65,93] All 264 descriptors were used (see **Supplementary Computational Methods** and **Data S3** for full list of descriptors). Algorithms tested include Random Forest (RF), K-Nearest Neighbors (KNN), Classification and Regression Tree (CART), Linear Discriminant Analysis (LDA), and Support Vector Machine (SVM). (B) Recursive feature elimination (RFE) with the number of features selected on the x-axis and mean test accuracy on the y-axis (Accuracy = (TP + TN) / (TP + TN + FP + FN)). (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 mode's classification performance, where the rows represent the actual classes (high- or low-reactive) and the columns represent the predicted classes. (D) SHapley Additive exPlanations (SHAP) summary showing the impact of selected features on the predicted classification (high- or low-reactive cysteines).[62,68] 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.

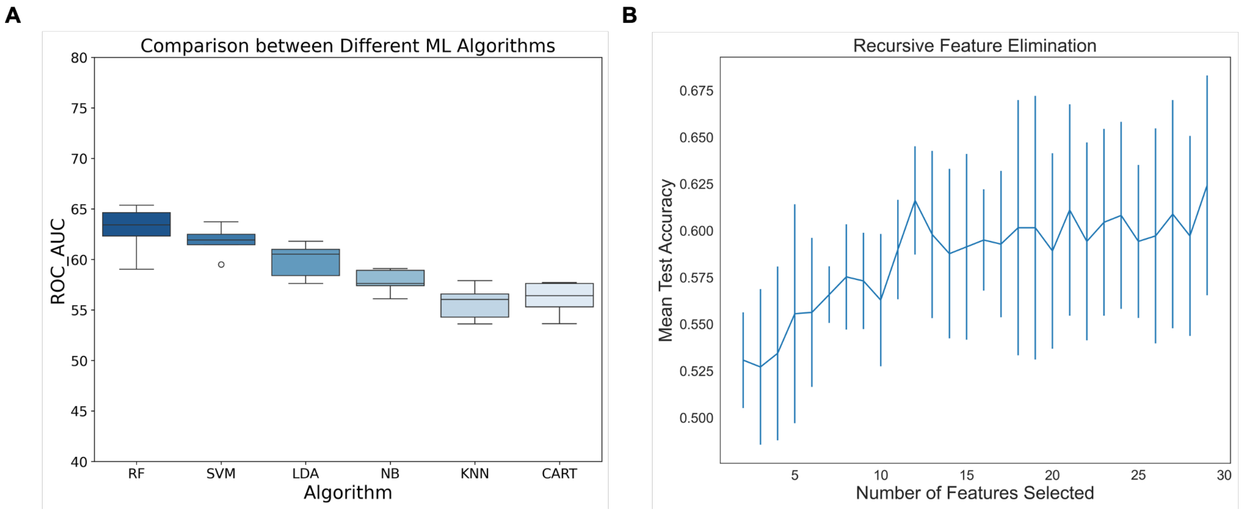


Figure S16. (A) Algorithm comparison of training data after training on experimental dataset and testing on aggregated dataset from the literature. Algorithms tested include Random Forest (RF), K-Nearest Neighbors (KNN), Classification and Regression Tree (CART), Linear Discriminant Analysis (LDA), and Support Vector Machine (SVM). (B) Recursive feature elimination (RFE) with the number of features selected on the x-axis and mean test accuracy on the y-axis (Accuracy = (TP + TN) / (TP + TN + FP + FN)). The optimum number of features chosen is 29.

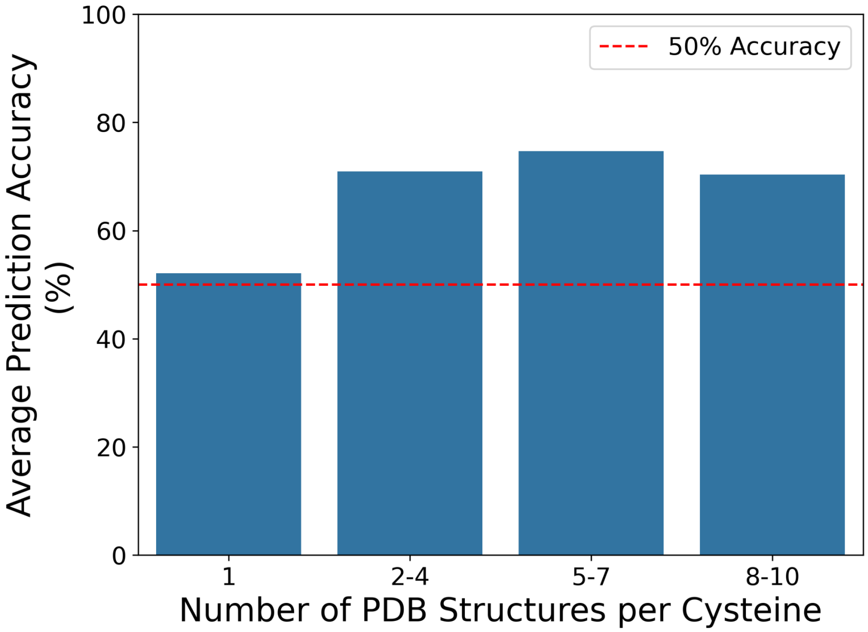


Figure S17. (A) Bar chart comparing the average prediction accuracy for test set cysteines with multiple PDB structures. For each cysteine, the prediction accuracy was calculated as the number of correct predictions divided by the total number of predictions. Cysteines were binned into four groups based on the number of PDB structures used during testing (1, 2–4, 5–7, and 8–10). The average prediction accuracy for each group is displayed, with the red horizontal line representing 50% accuracy.

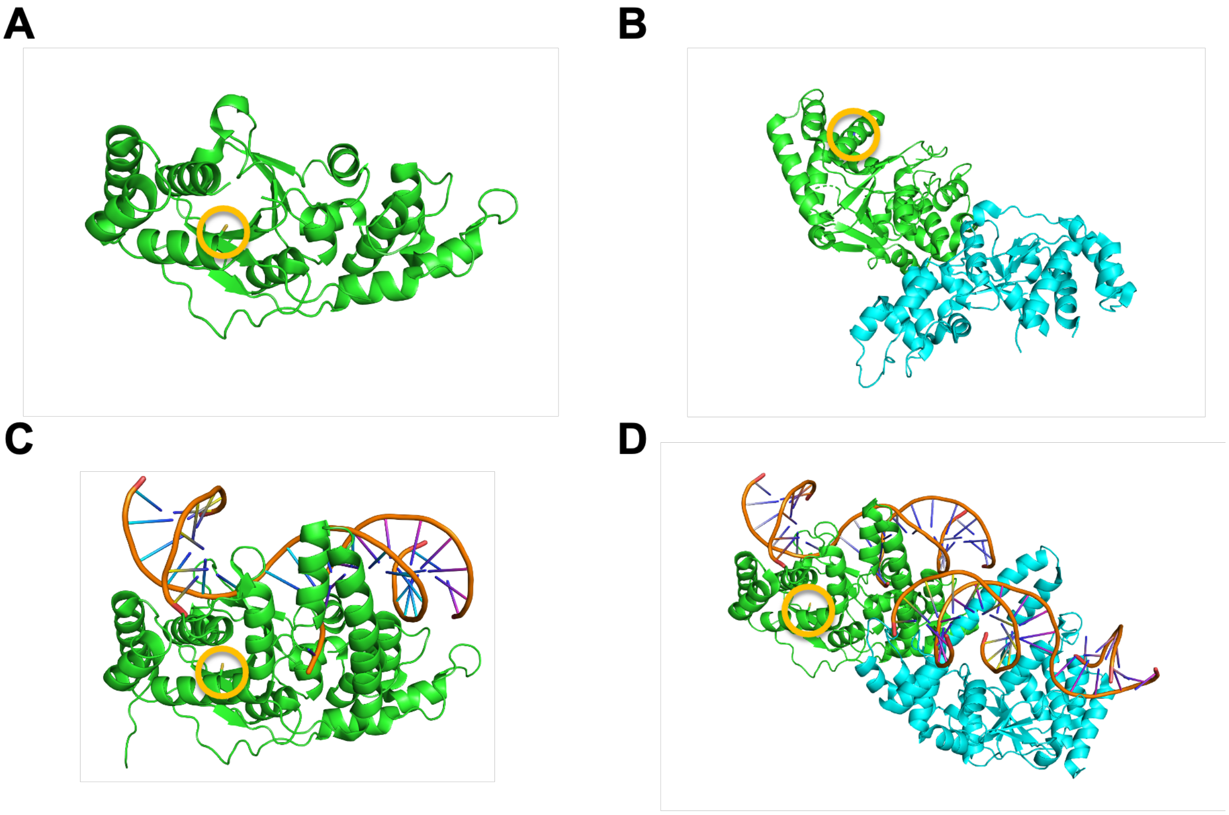


Figure S18. PDB structural representations utilized during testing for experimentally determined high-reactive cysteine C141 of Flap Endonuclease (FEN1). PDBs 5ZOD, 5FV7, 3Q8K, and 3Q8M were aligned, and the structures are displayed separately for clarity. Cysteine 141 is highlighted in yellow, with chain A shown in green and chain B in blue. (A) correct prediction of high-reactive C141 using PDB 5ZOD-with only chain A resolved. (B) correct prediction of high-reactive C141 using PDB 5FV7-with chains A and B resolved. (C) correct prediction of high-reactive C141 using PDB 3Q8K-with chain A resolved, and DNA bound. (D) incorrect prediction of high-reactive C141 using PDB 3Q8M-with chains A and B resolved, as well as having DNA bound.

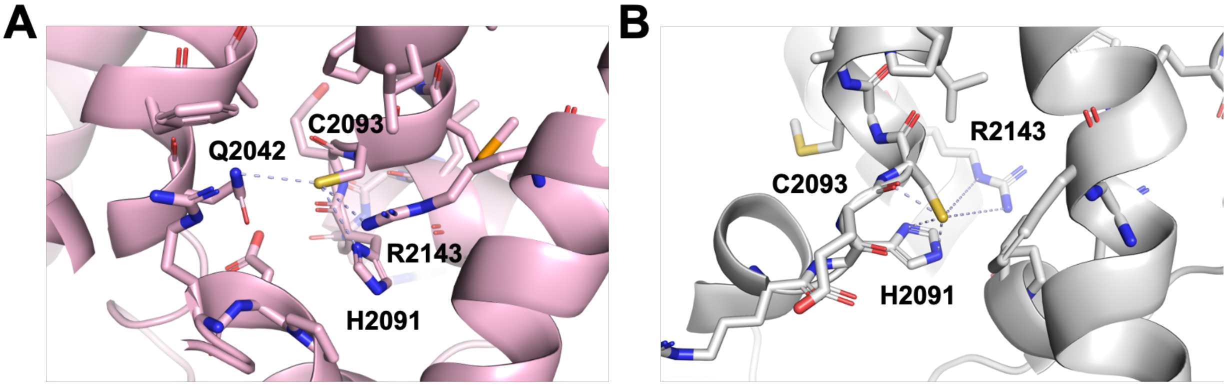


Figure S19. PDB structural representations utilized during testing for experimentally determined high-reactive cysteine C2093 of DNA-dependent protein kinase (DNA-PK). Local microenvironments of C2093 are depicted in the inactive Apo-DNA-PKcs conformation (A) and active DNA-PK holoenzyme-with DNA, Ku70 and Ku80 are resolved. C2093 is labeled, as well as neighboring hydrogen bond partners. (A) Incorrect prediction of high-reactive C2093 using PDB 5LUQ-the inactive Apo-DNA-PKcs conformation. (B) Correct prediction of high-reactive C2093 using PDB 5Y3R-the active DNA-PK holoenzyme.

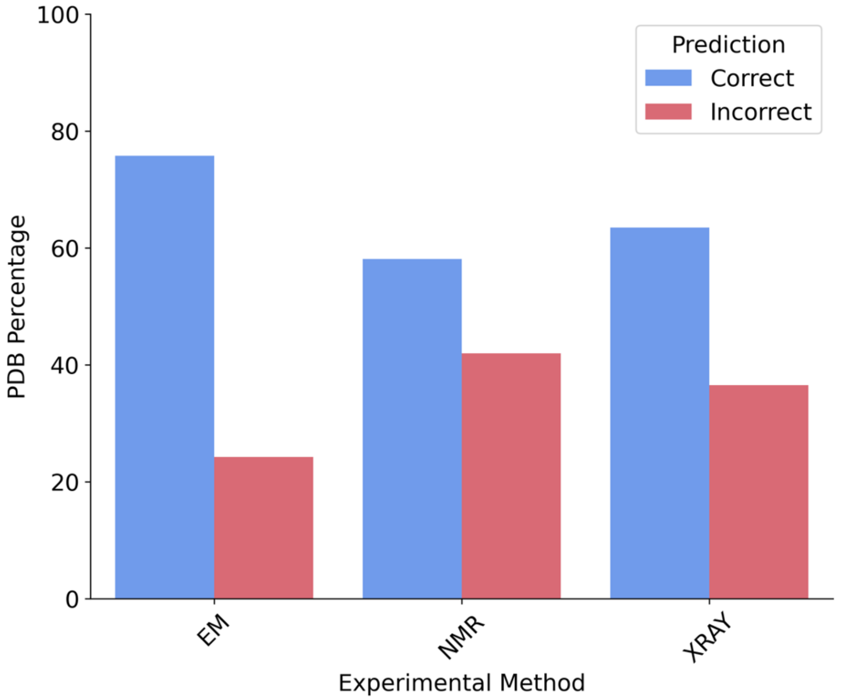


Figure S20. (A) Bar chart comparing the overall percentage of cases predicted correctly (blue) and incorrectly (red) by experimental structure determination methods. See **Data S3**.

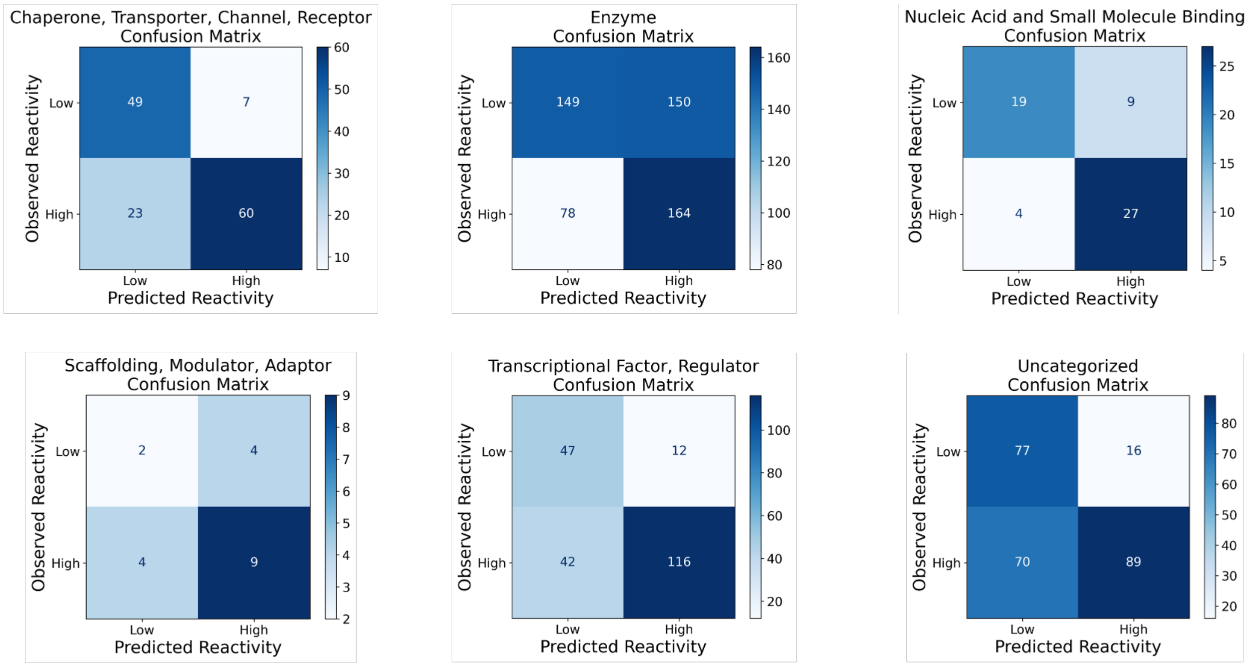


Figure S21. Confusion matrix heat maps for six protein classes showing the distribution of true positive, false positive, true negative, and false negative cases from the random forest algorithm for each protein class. The matrix provides a visual representation of the model's classification performance, where the rows represent the actual classes (high- or low-reactive) and the columns represent the predicted classes. The observed reactivity classes are based on quantitative cysteine reactivity isoTOP-ABPP ratios ($R_{10:1}$).

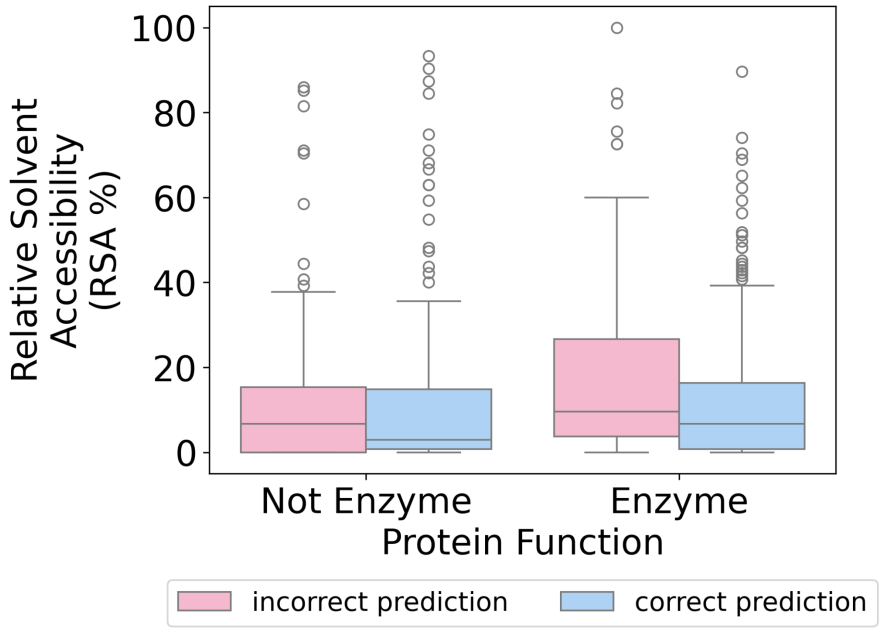


Figure S22. There is a relationship between relative solvent accessibility (RSA, %) and enzymatic function in proteins within the test set (Mann-Whitney U Test statistic: 48,679, p = 0.0056).

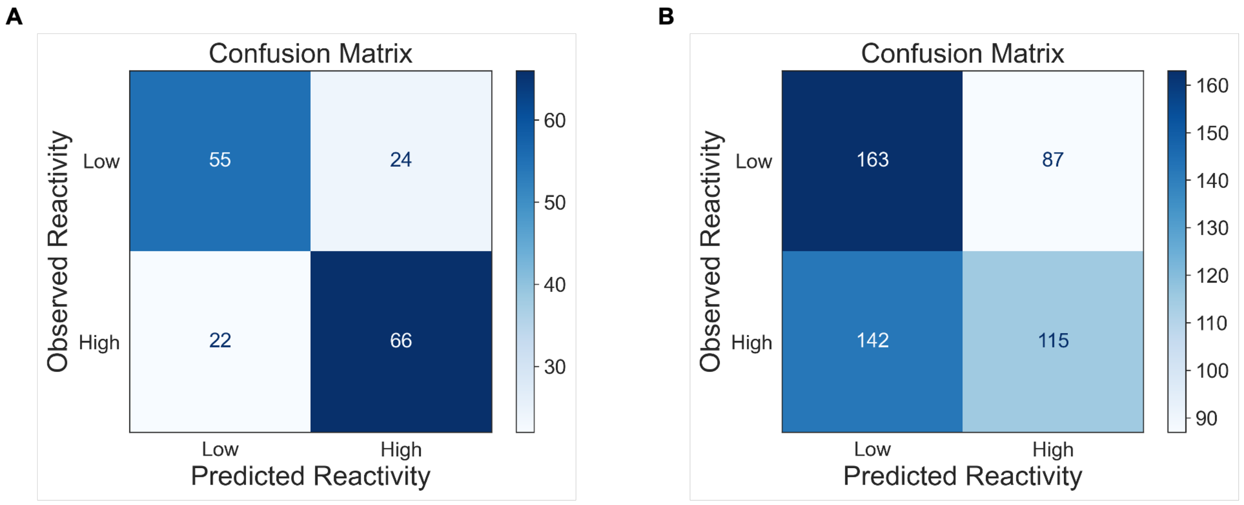


Figure S23. (A) For training cases that were correctly predicted using PDB structures, descriptors were generated for the same cysteine identifiers but using their corresponding AlphaFold 2 structures. Confusion matrix heatmap showing the distribution of true positive, false positive, true negative, and false negative cases from the random forest algorithm for each protein class. The matrix provides a visual representation of the model's classification performance, where the rows represent the actual classes (high- or low-reactive) and the columns represent the predicted classes. The observed reactivity classes are based on quantitative cysteine reactivity isoTOP-ABPP ratios ($R_{10:1}$). (B) For cysteines identified in our experimental dataset (**Data S1**) that did not have an associated PDB structure or did not have their cysteine resolved in an associated PDB structure. Descriptors were generated for these cysteine identifiers but using their corresponding AlphaFold 2 structures. Confusion matrix heatmap showing the distribution of true positive, false positive, true negative, and false negative cases from the random forest algorithm for each protein class. The matrix provides a visual representation of the model's classification performance, where the rows represent the actual classes (high- or low-reactive) and the columns represent the predicted classes. The observed reactivity classes are based on quantitative cysteine reactivity isoTOP-ABPP ratios ($R_{10:1}$).

3.5.2 - Supplementary Tables

Table 3.1 Files in Proteomics Identification Database[94,95] (PRIDE) datasets. PRIDE IDENTIFIER: PXD056064.

Figure	File name	Experiment
Figure 1	2022-03-23-KB-LB-10v100-Iso-1	isoTOP
	2022-03-23-KB-LB-10v100-Iso-2	isoTOP
	2022-03-23-KB-LB-10v100-Iso-3	isoTOP
	2022-03-23-KB-LB-10v100-Iso-4	isoTOP
	2022-03-23-KB-LB-10v100-Iso-5	isoTOP
	2022-03-23-KB-LB-10v100-Iso-6	isoTOP
	2022-03-23-KB-LB-10v100-Iso-7	isoTOP
	2022-03-23-KB-LB-10v100-Iso-8	isoTOP
	2022-03-30-KB-LB-10v100-Iso-9	isoTOP
	2022-03-30-KB-LB-10v100-Iso-10	isoTOP
	2022-03-30-KB-LB-10v100-Iso-11	isoTOP
	2022-03-30-KB-LB-10v100-Iso-12	isoTOP
	2022-03-30-KB-LB-10v100-Iso-13	isoTOP

Table 3.2 Experimental case studies to investigate the relationship between cysteine high-reactivity and experimentally determined pKa.

Identifier	Gene	Reference	Experimental pKa	isoTOP-ABPP Ratio
P16455_C145	MGMT	Guengerich et al. 2003	5.3	0.88
P16455_C150	MGMT	Guengerich et al. 2023	NA	1.33
Q99497_C46	PARK7	Witt et al. 2008	NA	5.1
Q99497_C53	PARK7	Witt et al. 2008	NA	6.3
Q99497_C106	PARK7	Witt et al. 2008	5.4	2.13
P10599_C32	TXN	Forman-Kay et al. 1992	6.3	1.37
P10599_C35	TXN	Forman-Kay et al. 1992	NA	1.65
P10599_C73	TXN	Forman-Kay et al. 1992	NA	5.06
P07237_C36	TXN	Conway et al.	4.5	NA

		2015		
P07237_C53	TXN	Conway et al. 2015	NA	1.38
P07237_C312	TXN	Conway et al. 2015	NA	6.39
P07237_C343	TXN	Conway et al. 2015	NA	4.72
P07237_C400	TXN	Conway et al. 2015	NA	1.15
P63146_C88	UBE2B	Tolbert et al. 2005	10.2	6.04

3.6 - 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.

79. Yu, F., Teo, G.C., Kong, A.T., Haynes, S.E., Avtonomov, D.M., Geiszler, D.J., and Nesvizhskii, A.I. (2020). Identification of modified peptides using localization-aware open search. Nat. Commun. *11*, 4065. 10.1038/s41467-020-17921-y.

80. Teo, G.C., Polasky, D.A., Yu, F., and Nesvizhskii, A.I. (2021). Fast deisotoping algorithm and its implementation in the msfragger search engine. J. Proteome Res. *20*, 498–505. 10.1021/acs.jproteome.0c00544.

81. Shteynberg, D.D., Deutsch, E.W., Campbell, D.S., Hoopmann, M.R., Kusebauch, U., Lee, D., Mendoza, L., Midha, M.K., Sun, Z., Whetton, A.D., et al. (2019). PTMProphet: Fast and Accurate Mass Modification Localization for the Trans-Proteomic Pipeline. J. Proteome Res. *18*, 4262–4272. 10.1021/acs.jproteome.9b00205.

82. Rose, P.W., Prlić, A., Altunkaya, A., Bi, C., Bradley, A.R., Christie, C.H., Costanzo, L.D., Duarte, J.M., Dutta, S., Feng, Z., et al. (2017). The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Res. *45*, D271–D281. 10.1093/nar/gkw1000.

83. Zhang, S., Krieger, J.M., Zhang, Y., Kaya, C., Kaynak, B., Mikulska-Ruminska, K., Doruker, P., Li, H., and Bahar, I. (2021). ProDy 2.0: increased scale and scope after 10 years of protein dynamics modelling with Python. Bioinformatics *37*, 3657–3659.

10.1093/bioinformatics/btab187.

84. Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., et al. (2010). Towards a knowledge-based Human Protein Atlas. Nat. Biotechnol. *28*, 1248–1250. 10.1038/nbt1210-1248.

85. Han, X., Wang, J., Wang, J., Liu, S., Hu, J., Zhu, H., and Qian, J. (2017). ScaPD: a database for human scaffold proteins. BMC Bioinformatics *18*, 386. 10.1186/s12859-017-1806-6.

86. Thompson, J.D., Gibson, T.J., and Higgins, D.G. (2002). Multiple sequence alignment using ClustalW and ClustalX. Curr. Protoc. Bioinformatics *Chapter 2*, Unit 2.3. 10.1002/0471250953.bi0203s00.

87. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

88. McKinney, W. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference Proceedings of the python in science conference. (SciPy), pp. 56–61. 10.25080/Majora-92bf1922-00a.

89. Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. *9*, 90–95. 10.1109/MCSE.2007.55.

90. Waskom, M. (2021). seaborn: statistical data visualization. JOSS *6*, 3021. 10.21105/joss.03021.

91. Lundberg, S., and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. arXiv. 10.48550/arxiv.1705.07874.

92. UniProt Consortium (2023). Uniprot: the universal protein knowledgebase in 2023. Nucleic Acids Res. *51*, D523–D531. 10.1093/nar/gkac1052.

93. Castellón, J.O., Ofori, S., Armenta, E., Burton, N., Boatner, L.M., Takayoshi, E.E., Faragalla, M., Zhou, A., Tran, K., Shek, J., et al. (2023). Chemoproteomics identifies proteoform-selective caspase-2 inhibitors. BioRxiv. 10.1101/2023.10.25.563785.

94. Deutsch, E.W., Csordas, A., Sun, Z., Jarnuczak, A., Perez-Riverol, Y., Ternent, T., Campbell, D.S., Bernal-Llinares, M., Okuda, S., Kawano, S., et al. (2017). The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. Nucleic Acids Res. *45*, D1100–D1106. 10.1093/nar/gkw936.

95. Perez-Riverol, Y., Bai, J., Bandla, C., García-Seisdedos, D., Hewapathirana, S., Kamatchinathan, S., Kundu, D.J., Prakash, A., Frericks-Zipper, A., Eisenacher, M., et al. (2022). The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. Nucleic Acids Res. *50*, D543–D552. 10.1093/nar/gkab1038.

Chapter 4: Conclusion

This dissertation aims to advance our understanding of the therapeutic accessibility of the human cysteinome through the integration of experimental chemoproteomics and computational modeling. By addressing key limitations in coverage, data integration, and mechanistic insight, this work provides foundational tools for studying cysteine reactivity and druggability, enabling future innovations in precision medicine and covalent drug discovery.

In Chapter 1, we introduced MS-CpDAA (Mass Spectrometry-based Chemoproteomics Detected Amino Acid Analysis Suite; https://github.com/lmboat/ms_cpdaa_analysis), an automation software designed to streamline residue-level aggregation and data analysis for high-throughput chemoproteomics experiments. Leveraging MS-CpDAA, we quantified the performance of a novel workflow combining single-pot, solid-phase-enhanced sample preparation with high-field asymmetric waveform ion mobility spectrometry (SP3-FAIMS), enabling high-coverage profiling of the cysteinome.[1] This approach expanded coverage to 13% of the cysteinome (34,225 cysteines)—the highest achieved within our group—by integrating optimized workflows and advanced data analysis pipelines. Beyond profiling, MS-CpDAA supported applications such as multiplexed CuAAC Suzuki-Miyaura chemoproteomics (mCSCP)[2] and the evaluation of Tunable Amine-Reactive Electrophiles (TARE probes),[3] showcasing its versatility and scalability for chemoproteomic research.

Building on these advancements, Chapter 2 established CysDB, a repository of human cysteine chemoproteomics data derived from nine high-coverage studies.[4] CysDB is an SQL database accessible via an accompanying web application (https://backuslab.shinyapps.io/cysdb/). It features chemoproteomic measures of identification, ligandability, and hyper-reactivity for 62,888 cysteines (24% of the cysteinome), along with structural and functional annotations from resources like UniProtKB/Swiss-Prot, Cancer Gene Census, ClinVar, and the Protein Data Bank (PDB).

To address the lack of standardized approaches for comparing cysteine chemoproteomic datasets, CysDB supports inter- and intra-dataset analyses and provides tools for prioritizing cysteine residues based on reactivity, ligandability, and protein-level druggability. This resource bridges experimental measures with functional insights, enabling both broad-scale analyses of the cysteinome and targeted investigations of therapeutically relevant proteins. Most importantly, CysDB was designed to incorporate new datasets to further support the continued growth of the druggable cysteinome. CysDB has been updated since its initial release. V1.5 now includes data for 64,681 cysteines, representing 25% of the cysteinome, along with metrics of redox sensitivity[5,6] and high-throughput screening results,[7] such as the identification of a novel pan-cysteine reactive chemotype, phenylpropiolate (PP). These enhancements demonstrate CysDB's adaptability and its value as a dynamic resource for advancing cysteine-targeted research.

Chapter 3 leveraged the data and structural insights from CysDB to investigate cysteine hyper-reactivity through computational modeling and experimental expansion. To enhance the training dataset, we conducted 13 new experiments, identifying 640 hyper-reactive cysteines not previously documented in CysDB. This effort doubled the number of known hyper-reactive cysteines, enriching the dataset for predictive modeling and enabling a more comprehensive comparison with existing entries in CysDB. Initial analysis focused on primary sequence descriptors to identify patterns or trends in cysteine reactivity. However, not a single feature alone was sufficient to describe hyper-reactivity toward IAA, prompting a shift to three-dimensional structural analysis.

Using CysDB's structure-mapping pipeline, we examined 3D features, including residue proximity, solvent accessibility, secondary structure, and predicted pKa, to identify characteristics associated with hyper-reactivity. We also analyzed features linked to ligandable cysteines, such as those reactive with acrylamides and chloroacetamides, but found that none of these descriptors alone were sufficient to predict hyper-reactivity to IAA. This highlights the distinct and multifactorial nature of IAA reactivity compared to ligandability.

To address this complexity, we developed CIAA (Cysteine reactivity towards IodoAcetamide Alkyne), a structure-guided computational model that integrates primary sequence and 3D structural features into a random forest algorithm. External validation of CIAA revealed key structural drivers of cysteine reactivity, such as backbone hydrogen bond donor atoms, and highlighted gaps in current modeling approaches, including challenges with protein structure selection and dataset curation. These findings emphasize how artificial intelligence can provide mechanistic insights into cysteine reactivity, offering tools to inform the design of covalent inhibitors and advance cysteine-targeted drug discovery.

MS-CpDAA expedited the identification of sites of covalent modification in high-throughput chemoproteomics experiments, facilitating the aggregation of cysteine reactivity data into CysDB. CysDB, in turn, serves as a resource for integrating and prioritizing chemoproteomic data across datasets. CIAA leveraged CysDB’s structural annotation pipeline to uncover structural drivers of hyper-reactivity and provide mechanistic insights into cysteine reactivity. These tools collectively address challenges in data analysis, integration, and mechanistic understanding, significantly expanding the scope of cysteine-targeting strategies.

While this dissertation focuses on cysteines, the methodologies developed are readily adaptable to other reactive residues, such as lysines and tyrosines, and can be applied to explore additional therapeutic spaces. Future directions include expanding structural datasets, incorporating dynamic protein conformations, and refining predictive algorithms to enhance modeling accuracy and capture the complexity of protein-ligand interactions. By bridging experimental chemoproteomics with computational modeling, this work advances the therapeutic accessibility of cysteine residues, laying a foundation for leveraging the cysteinome to address unmet therapeutic needs in precision medicine and covalent drug design.

4.1 - References

1. 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.

2. 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. 10.1021/acs.analchem.0c04726.

3. Tang, K.-C., Cao, J., Boatner, L.M., Li, L., Farhi, J., Houk, K.N., Spangle, J., Backus, K.M., and Raj, M. (2022). Tunable Amine-Reactive Electrophiles for Selective Profiling of Lysine. Angew. Chem. Int. Ed *61*, e202112107. 10.1002/anie.202112107.

4. 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.

5. Desai, H.S., Yan, T., Yu, F., Sun, A.W., Villanueva, M., Nesvizhskii, A.I., and Backus, K.M. (2022). SP3-Enabled Rapid and High Coverage Chemoproteomic Identification of Cell-State-Dependent Redox-Sensitive Cysteines. Mol. Cell. Proteomics *21*, 100218. 10.1016/j.mcpro.2022.100218.

6. Yan, T., Julio, A.R., Villanueva, M., Jones, A.E., Ball, A.B., Boatner, L.M., Turmon, A.C., Nguyễn, K.B., Yen, S.L., Desai, H.S., et al. (2023). Proximity-labeling chemoproteomics defines the subcellular cysteinome and inflammation-responsive mitochondrial redoxome. Cell Chem. Biol. *30*, 811-827.e7. 10.1016/j.chembiol.2023.06.008.

7. 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.