The Activation Score: Predicting cysteine reactivity from chemoproteomics data and protein language model (PLM) embeddings

Emily Lachtara, Yu-Hsin Chao, Karsten Krug, Johannes Hermann, Rohan Varma

Frontier Medicines Corporation, Boston, MA and South San Francisco, CA

Background

- Covalent drugs have been used to treat disease for over a century (1), but the streamlined discovery and design of covalent drugs starting from a fragment is a new and promising field. Covalent drugs bind irreversibly or reversibly and can affect a protein's activity or stability. The covalent approach provides an advantage when targeting "undruggable" proteins with poorly defined pockets or disordered domains, providing an anchor point to build potent and selective drugs.
- Due to its pKa range, cysteine is the preferred residue for covalent modification, thus there is great interest in targeting these residues for drug-discovery. Chemoproteomics is an approach to understand the interaction between small molecules and proteins in a cellular context. The Frontier™ Platform is built using chemoproteomics to enable the rapid and proteome-wide discovery of druggable cysteines.
- Here we present Frontier's Activation Score: a machine learning (ML)-based score derived from training on Frontier's vast chemoproteomics dataset. The Activation Score enables us to rank-prioritize chemically reactive cysteines across the proteome informing drug discovery strategy and accelerating covalent drug discovery.

Frontier Medicine's Druggability Atlas™ has identified >150,000 cysteine hotspots for covalent modification enabling access to >90% of the human proteome for drug discovery

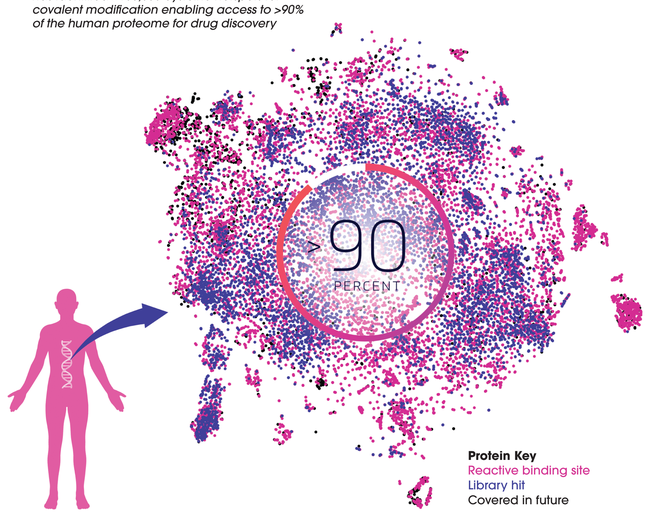


Figure 1: A TSNE projection was created from the canonical human proteome using PLM embeddings. Each dot represents a protein: Frontier's coverage of the human proteome is highlighted in pink and purple, whereas black dots are proteins that could be covered in the future.

Harnessing chemoproteomics to characterize reactive cysteines

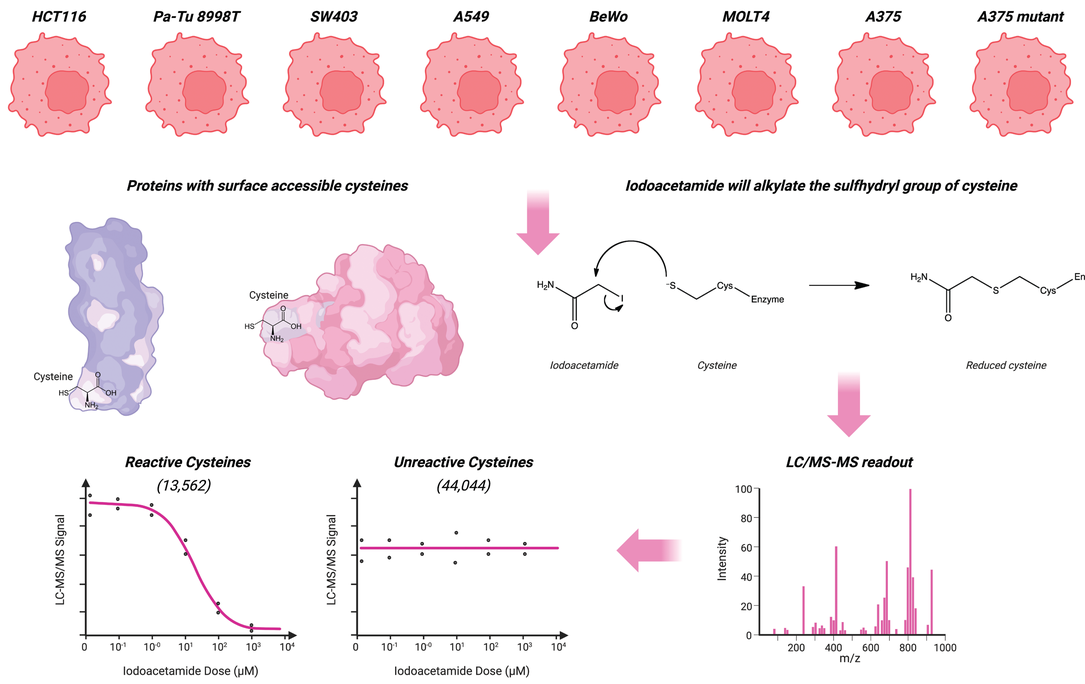


Figure 2: Eight cancer cell lines were treated with sulfhydryl-reactive alkylating Iodoacetamide (IAA) at 7 different concentrations (ranging from 0 to 1,000 μM) to block reduced cysteine residues for peptide mapping. A competition reaction between IAA and a cysteine-binding desthiobiotin Iodoacetamide (DBIA) chemical probe enabled enrichment of DBIA-labeled peptides in a dose-responsive manner. The samples were prepped and run on a mass spectrometer (MS) where loss of peptide signal indicated successful IAA-competition. We fit a four-parameter log-logistic model to the dose response data for each cysteine. When the model converged on a solution (there was a dose response) we labelled the cysteine as reactive, when there was no dose response, we labelled it as unreactive.

Protein language model embeddings capture important biophysical, structural, and functional properties of residues

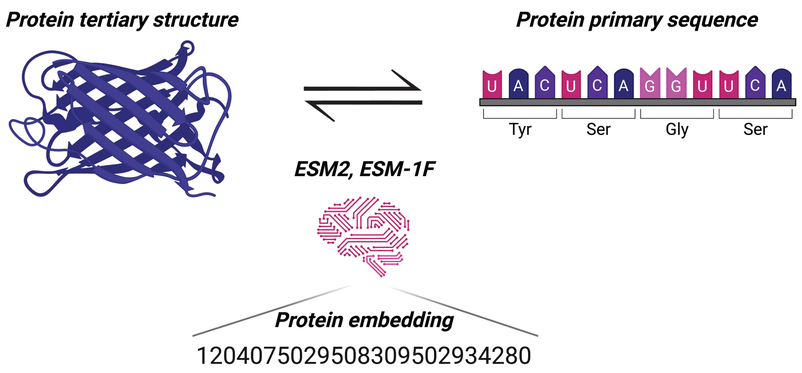


Figure 3: Frontier's proprietary language model incorporates an evolutionary scale model (ESM2) and an inverse folding model (ESM-IF1) (2,3). Evolutionary Scale Modeling (ESM2) is a type of PLM that uses uniref50 sequence DB as training data. ESM-IF1 is an inverse folding model that utilizes Alphafold protein structures to predict the protein sequence from its backbone atom coordinates. The model produces an embedding which is a pretrained descriptor that simultaneously takes as input both primary and tertiary protein structure, thus encoding both structural (Alphafold) and sequential information about the protein. We applied the model to Frontier's Druggability Atlas™ and generated embeddings from the canonical human proteome.

The Activation Score model combines chemoproteomics with protein language model embeddings to predict cysteine reactivity

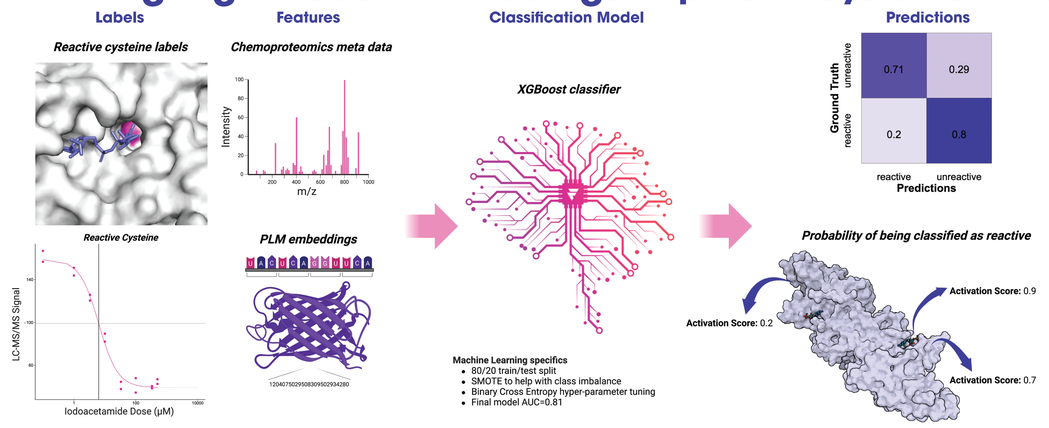


Figure 4: We trained several different ML models to classify reactive from unreactive cysteines using a feature set that included embeddings derived from a residue-level PLM and metadata from proprietary MS/MS spectra data derived from conducting thousands of isoTOP-ABPP (4) experiments. These MS/MS spectra data provide us with measures of reproducibility of modification, and abundance of peptide which can be used in our model. Among the tested ML models, we found that XGBoost had the highest performance classifying reactive from unreactive cysteines (AUC = 0.81). The Activation Score is the probability that a cysteine is classified as reactive by our machine learning model.

The Activation Score is a validated and powerful tool for prioritizing cysteines for covalent drug discovery

Figure 5: We applied our model to Frontier's hotspot database to calculate Activation Scores across the human proteome. To evaluate the model quantitatively, we partitioned the cysteines into quartiles utilizing the Activation Score: Q4 represents cysteines with the highest scores (> 0.75), and Q1 represents cysteines with the lowest scores (< 0.2). We utilized our chemoproteomic proteome-wide covalent library profiling to validate whether cysteines with high Activation Scores also demonstrated experimental reactivities. We also compared the Activation Score to physicochemical properties that confer reactivity (SASA, and pKa)

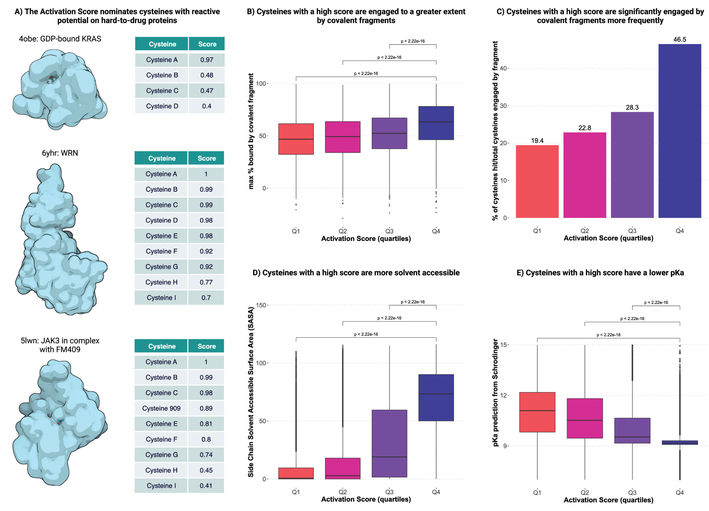
a. Pictured are the crystal structures of one drugged (JAK3) and two hard-to-drug (KRAS and WRN) proteins. Displayed are cysteines from Frontier's Druggability Atlas™ and their associated Activation Scores.

b. Boxplot depicting the max % bound by a covalent fragment for each cysteine derived from our proteome-wide covalent library profiling experiments. Percent bound is calculated from the competition of covalent fragments and DMSO (5); the more potent reactions will have a higher % bound because the covalent fragment outcompetes DMSO for the cysteine binding site.

c. Barchart displaying the frequency of significant covalent fragment-cysteine reactions (hits) compared to the total number of covalent fragment-cysteine reactions.

d. Boxplot depicting the distribution of cysteine's Solvent Accessible Surface Area (SASA). A greater SASA means the cysteine is more surface accessible.

e. Boxplot depicting the distribution of cysteine's predicted pKa (acid dissociation constant). The pKa predictions were obtained from Schrodinger Epik (6), where a lower pKa denotes a cysteine that is more susceptible to oxidation.



Methods

- IAA dose response samples were prepped for MS analysis utilizing Tandem Mass Tag (TMT) multiplexed quantification and then run on a Thermo Orbitrap Eclipse mass spectrometer. The resulting MS data were searched using the open-source Comet algorithm (7) using the informatics pipeline described in (5).
- We imported the ESM2 and ESM1 models from the esm python package (8). We then fed the models the pdb structures from the canonical human proteome Alphafold V2 release (9) to calculate embeddings for every protein.
- We tested the performance of XGBoost, Neural Net, Logistic Regression, Naïve Bayes, Random Forest, Decision Tree, and K-Nearest Neighbors classification models from the xgboost (10), tensorflow (11), and scikit-learn (12) packages in Python. We utilized Synthetic Minority Oversampling Technique (SMOTE) from the imblearn package (13) to correct class imbalance within our training set.
- The figures were created with BioRender https://www.biorender.com/

Conclusions

- We developed the Activation Score, a ML/AI-based model to rank-prioritize chemically reactive cysteines within a protein or across the proteome.
- Cysteine reactivity can be predicted by leveraging chemoproteomics data with PLM embeddings in an algorithmic approach.
- The Activation Score is highly correlated with physicochemical properties such as SASA and pKa, but key differences exist.
- Cysteines with a higher Activation Score demonstrate more frequent and more notable engagement in Frontier's proteome wide covalent library profiling experiments.
- The Activation Score nominates cysteines with reactive potential across the proteome including proteins that are hard to drug but important for disease (such KRAS and WRN).
- The Activation Score is incorporated into Frontier's Druggability Atlas™ and presents a powerful tool for prioritizing cysteines for covalent drug discovery.

References

1. Boike, L., Henning, N. J., & Nomura, D. K. (2022). Advances in covalent drug discovery. *Nature reviews. Drug discovery*, 21(12), 881–898.
2. Rives et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 118(15).
3. Hsu et al. (2022). Learning inverse folding from millions of predicted structures. *Proceedings of the 39th International Conference on Machine Learning*, PMLR 162:8946-8970.
4. Weerapana, E., Speers, A. E., & Cravatt, B. F. (2007). Tandem orthogonal proteolysis-activity-based protein profiling (TOP-ABPP)–a general method for mapping sites of probe modification in proteomes. *Nature protocols*, 2(6), 1414–1425.
5. Kuljanin et al. (2021). Reimagining high-throughput profiling of reactive cysteines for cell-based screening of large electrophile libraries. *Nature biotechnology*, 39(5), 630–641.
6. Johnston et al. (2023). Epik: pK_a and Protonation State Prediction through Machine Learning. *Journal of chemical theory and computation*, 19(8), 2380–2388.
7. Eng, J. K., Jahan, T. A., & Hoopmann, M. R. (2013). Comet: an open-source MS/MS sequence database search tool. *Proteomics*, 13(1), 22–24.
8. Barbi et al. (2021). ESM-Tools version 5.0: a modular infrastructure for stand-alone and coupled Earth system modelling (ESM). *Geoscientific Model Development*, 14, 4051-4067.
9. Jumper et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589.
10. Chen et al. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).
11. Abadi et al. (2016). Tensorflow: A system for large-scale machine learning. In *Proceeedings of the 12th USENIX Symposium on Operating Systems Design and Implementation* (pp. 265–283).
12. Pedregosa et al.(2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct), 2825–2830.
13. Lemaitre, G., Nogueira, F., & Aridas, C.K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*,18, 1-5.