# CETSA Feature Based Clustering for Protein Outlier Discovery by Protein-to-Protein Interaction Prediction

Xulei Yang[1*], Qing Da[2*], Peisheng Qian[1*], Bharadwaj Veeravalli[2], Tam Wai Leong[3,4], Lingyun Dai[5,6], Pär Nordlund[6,7], Nayana Prabhu[6], Ziyuan Zhao[1] and Zeng Zeng[1†]

***Abstract*— The Cellular Thermal Shift Assay (CETSA) is a biophysical assay based on the principle of ligand-induced thermal stabilization of target proteins. This technology has revolutionized cell-based target engagement studies and has been used as guidance for drug design. Although many applications of CETSA data have been explored, the correlations between CETSA data and protein-protein interactions (PPI) have barely been touched. In this study, we conduct the first exploration study applying CETSA data for PPI prediction. We use a machine learning method, Decision Tree, to predict PPI scores using proteins' CETSA features. It shows promising results that the predicted PPI scores closely match the ground-truth PPI scores. Furthermore, for a small number of protein pairs, whose PPI score predictions mismatch the ground truth, we use iterative clustering strategy to gradually reduce the number of these pairs. At the end of iterative clustering, the remaining protein pairs may have some unusual properties and are of scientific value for further biological investigation. Our study has demonstrated that PPI is a brand-new application of CETSA data. At the same time, it also manifests that CETSA data can be used as a new data source for PPI exploration study.**

## I. Introduction

Cellular Thermal Shift Assay (CETSA) [1] is a biophysical assay based on the principle of ligand-induced thermal stabilization of target proteins, meaning that a protein's melting temperature will change upon ligand interaction. CETSA's unique experimental approach allows the assessment and quantification of target engagement under physiological conditions – without the need to modify the compound or protein. This provides data that is both actionable and biologically relevant.

CETSA data has a wide range of applications, mainly in the field of drug discovery. It can be used in Library Screening [2], Translational studies [3], Target Deconvolution [4], Biomarker Discovery [5], Selectivity [6], etc. However, the correlations between CETSA data and protein-protein interactions (PPI) have barely been touched. In this study, we make an effort to conduct exploration study towards this direction. The CETSA dataset used in this study is taken from [7]. In this dataset, there are a total of 6 cell lines and we use HCT116 and HEK293T for our research. Each protein in CETSA dataset has 10 features, corresponding to its abundance under 37°C, 40°C, 43°C, 46°C, 49°C, 52°C, 55°C, 58°C, 61°C, and 64°C.

What we are trying to explore is applying the CETSA data to predict PPI scores. PPIs are physical contacts of high specificity established between two or more protein molecules, as a result of biochemical events steered by interactions that include electrostatic forces, hydrogen bonding and the hydrophobic effect. PPI are critical to the functioning of cells and biological processes of all organisms. Identification of protein interactions can lead to a better understanding of biochemical metabolism of proteins.

PPI score between two proteins is a normalized value between $[0, 1]$. The bigger score indicates a higher similarity between the two proteins. The dataset that we use for the ground truth of PPI is called Bioplex [8] (biophysical interactions of ORFeome-based complexes). It is an open access resource for studying protein-protein interactions. There are two cell lines [9] in Bioplex only, which are HCT116 and HEK293T. Since there are 6 cell lines in CETSA dataset, we choose the same two cell lines (HCT116 and HEK293T) in CETSA dataset for our study.

To the best of our knowledge, we are the first one to apply CETSA data to predict PPI scores. Based on machine learning techniques, our PPI predictions based on CETSA features of the protein pairs are closely matching the Bioplex PPI ground truth. Iterative clustering is further used to reduce the number of mismatched protein pairs and the final remaining "outlier" pairs may have scientific values to biologists.

The rest of this paper is organized as follows. Section II reviews the related experimental approaches for PPI predictions. Our proposed methodology, i.e., PPI prediction and iterative clustering, is presented in Section III. Section IV provides the experimental results, and the conclusion is given in Section V.

## II. Related Work

There are a number of experimental approaches for PPI predictions as shown in Fig. 1, such as Yeast two-hybrid screening [10], Affinity purification coupled to mass spectrometry [11][12], Nucleic acid programmable protein array (NAPPA) [13], Intragenic complementation [14] and many others [15]. Protein interactions are very complicated and each experimental method only measures some of protein's

This work is supported by Competitive Research Programme "NRF-CRP22-2019-0003", National Research Foundation Singapore. *Contributed equally, †Corresponding author, email: zengz@i2r.a-star.edu.sg. [1]Institute for Infocomm Research (I2R), Agency for Science, Technology and Research, Singapore. [2]National University of Singapore. [3]Genome Institute of Singapore (GIS), Agency for Science, Technology and Research (A*STAR), [4]Cancer Science Institute of Singapore, National University of Singapore. [5]The First Affiliated Hospital of Southern University of Science and Technology, Shenzhen, China. [6]Institute of Molecular and Cell Biology, Agency for Science, Technology and Research (A*STAR), Singapore. [7]Department of Oncology and Pathology, Karolinska Institutet, Stockholm, Sweden.

features from its own perspective. Each of the experimental approaches [16] has its own strengths and weaknesses, especially with regard to the sensitivity and specificity.

As for our approach, CETSA data is obtained under a more physiological relevant condition compared with other experimental approaches. CETSA is based on the premise that upon heating, a protein will unfold and aggregate at a given temperature. We think the thermal stability of protein is also a very important aspect of protein features and might be an important indicator for PPI predictions. With that thought in mind, we apply the CETSA dataset for PPI score predictions.

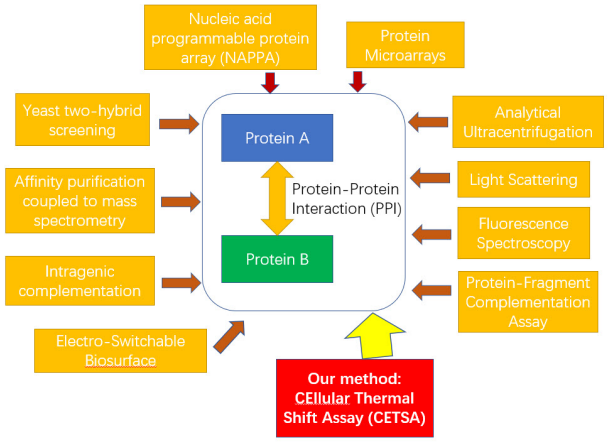


Fig. 1. Various experimental methods (as shown in yellow boxes) to predict PPI scores, our method (the red box) is under a more physiological relevant condition.

## III. Proposed Methodology

### A. PPI Prediction

Machine learning is a very useful tool in feature learning and feature prediction. Our goal is to train a machine learning model to predict PPI scores between a protein pair based on protein's features under 10 different temperatures, as is shown in Fig. 2. Specifically we use Decision Tree as our machine learning model.

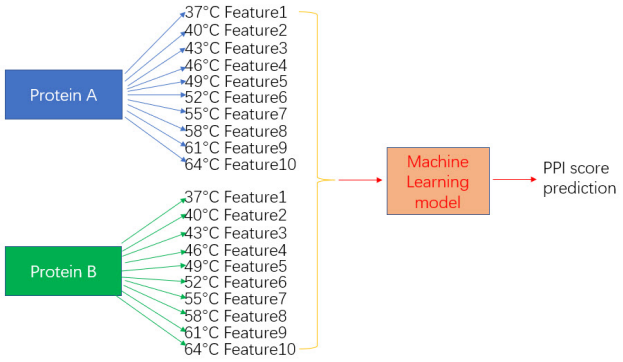


Fig. 2. Use machine learning model to predict PPI score for a protein pair based on their features under different temperatures

In order to train our Decision Tree model, we employ a 5-fold cross validation approach as shown in Fig. 3. We split all the protein pairs in the cell line into 5 folds with each fold containing 20% of the whole dataset. We take each fold as test set and the remaining folds as training set, then fit a model on the training set and evaluate it on the test set. In such as way, we obtain the predictions for the whole dataset.


Fig. 3. Diagram of 5-fold cross validation, we split all the protein pairs into 80% as training set and 20% as test set for each fold. Based on 5 predictions for each fold, we can have final PPI score prediction for all protein pairs in the cell line

After 5-fold validation, we have also tried different values of k for k-fold validation as well. The PPI prediction results of other values of k-fold also matches the PPI groundtruth, which further proves that the Decision Tree is a robust machine learning method for PPI prediction.

### B. Iterative Clustering

Our Decision Tree model has done well in PPI score prediction, but for a small number of protein pairs, their PPI score predictions do not match the ground truth. We think if these protein pairs are truly "mismatched", then they may be of great scientific values. So we take a further step to gradually reduces the number of these "mismatched" protein pairs by adopting Iterative Clustering Strategy to find the real "outliers".

The algorithm of iterative clustering is shown in Fig. 4. We can summarize our Iterative Clustering Algorithm as follows:

1. Use the Decision Tree model to fit the data of protein pairs. Feature importance may vary greatly based on the existing protein pairs.

2. Remove the protein pairs whose $MAE \leq 0.1$ (the green boxes in Fig. 4 in each iteration)

3. Keep the protein pairs whose $MAE > 0.1$ (the pink boxes in Fig. 4 in each iteration) for next iteration.

4. Repeat the process above, and output the final "outliers" (the pink box which is labelled as "others" in Fig. 4) at the end.

## IV. Experiments

### A. Dataset

The CETSA dataset used in this study is taken from [7], which consists of 6 cell lines (including HCT116 and HEK293T), and each protein contains 10 features from 10 temperatures.

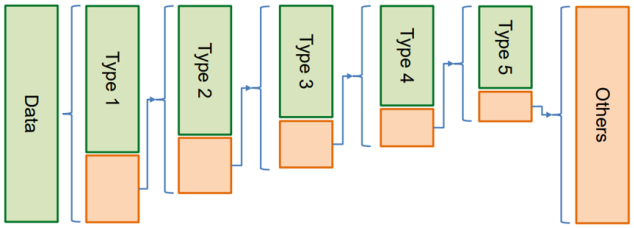


Fig. 4. Algorithm for iterative clustering, we gradually filter out the protein pairs with a small MAE (green boxes) and keep the proteins with a large MAE (pink boxes)

The ground truth PPI scores used in this study is taken from Bioplex, which consists of two cell lines only: HCT116 and HEK293T. We conduct the experiments on these two human cell lines. In CETSA dataset, there are a total of 7599 proteins in cell line HCT116 and 7945 proteins in cell line HEK293T. What we care about is protein-protein interaction, in which 2 proteins are involved. Based on the existing protein pairs in our ground truth dataset Bioplex, we can form 25485 protein pairs for cell line HCT116 and 41490 protein pairs for cell line HEK293T in CETSA dataset which have corresponding PPI scores in Bioplex dataset.

### B. Results of PPI Prediction

We employ a 5-fold cross validation scheme, and use Mean Square Error (MAE) as the performance indicator. Five-fold cross validation is conducted for PPI prediction by using the CETSA features from protein pairs in cell line HEK293T. The figure of PPI score ground-truths for the test set (which accounts for 20% of the total amount of protein pairs) in one fold is shown in the upper figure in Fig. 5 and PPI score predictions of the test set in this fold is shown in the bottom figure in Fig. 5. The shape of histograms for prediction and ground truth look alike, which indicates that the predicted PPI scores match the ground-truth PPI scores very well.

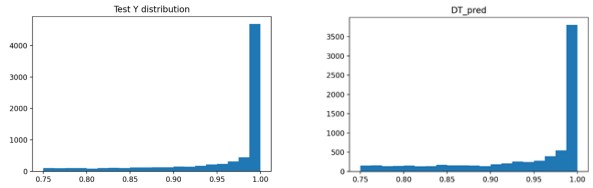


Fig. 5. PPI score ground-truths (upper figure) and predictions (bottom figure) of the test set in one fold in cell line HEK293T

Furthermore, we draw the distribution error between predictions and ground truth (for HEK293 cell line), as shown in Fig. 6. The average MAE is 0.0698 , the median MAE is 0.0456 and the max MAE is 0.2497, which indicates that our predictions are overall correct. This proves that our Decision Tree model can successfully predict PPI scores from CETSA features for most of the protein pairs.

Similar results can also be obtained using HCT116 cell line. Likewise, PPI predictions and PPI ground truths match each other and errors between predictions and ground truths remain low.

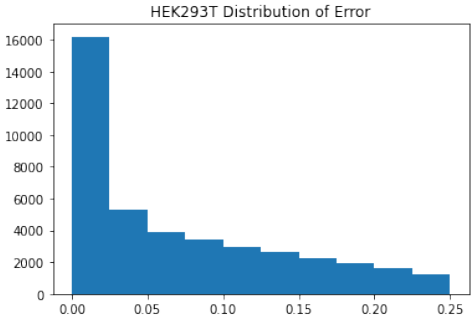


Fig. 6. PPI Error Distribution, the majority of protein pairs have MAE$<$ 0.05

### C. Results of Iterative Clustering

From Fig. 6, it can be seen that the MAE is low for most protein pairs, which means PPI prediction scores for most protein pairs match the ground truth quite well. However, there is still a small portion of protein pairs whose predictions does not match the ground truth well. This may due to some noise contained in CETSA features, or due to some kinds of "novel" property of CETSA features. The latter is worthy of further biological investigation. We therefore use iterative clustering to gradually reduce the number of mismatched protein pairs and filter out the final "outliers" which cannot be grouped into any of the clusters.

After each iteration, only protein pairs whose predicted PPI scores mismatch the ground truths by a large margin ($MAE > 0.1$) are selected as candidates for next iteration. In our study, we set the number of iterations to be 5. From Iteration 1 to Iteration 5, the total number of protein pairs in HEK293T that take part in each iteration drops significantly (41490 $\rightarrow$ 12656 $\rightarrow$ 5691 $\rightarrow$ 2926 $\rightarrow$ 1552 $\rightarrow$ 865). As is shown in Fig. 7, we can intuitively see the total number of protein pairs in each iteration (blue line) and the number of unmatched protein pairs after each iteration (purple line). The unmatched protein pairs will serve as data samples for next iteration.

Although most of the PPI score predictions from CESTA features gradually fit the ground truth during iterations, at the end of the 5th iteration, there are still 865 protein pairs (2.08% of the total protein pairs) whose PPI score predictions mismatch the ground-truths. These remaining protein pairs may have some unusual properties and are of great value for biologists for further study.

Same pattern can be found in HCT116 cell line as well. From Iteration 1 to Iteration 5, the total number of protein pairs that participate in each iteration decreases quickly (25485 $\rightarrow$ 7419 $\rightarrow$ 3146 $\rightarrow$ 1600 $\rightarrow$ 873 $\rightarrow$ 510). After the last iteration, 510 protein pairs (2.00% of the total protein pairs) still have PPI predictions that mismatch the ground-truths and they should be considered as final "outlier" protein pairs.

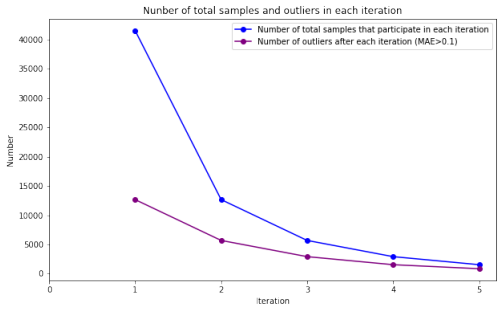


Fig. 7. Nunber of total samples (blue line) and outliers (purple line) in each iteration for HEK293T dataset.

In order to better fit the existing protein pairs, the importance of each protein feature changes during every iteration, which is shown in Fig. 8. Noted that for each protein pairs, there are 10 features corresponding to 10 temperatures. Feature 0 refers to the average value of features of 2 proteins under 37°C, Feature 1 $\rightarrow$ 40°C ,..., Feature 9 $\rightarrow$ 64°C. Feature 0 (37°C) is a constant and always equal to 0. We can see from Fig. 8 that:

1) Unlike using Euclidean Distance to measure the similarity between 2 proteins, where each feature shares the exact same weight, in our model, the importance of each feature is not equal to one another. And the distribution of feature importance varies greatly across iterations;

2) In the first two iterations, feature importance is stable, indicating such settings of feature importance are a good fit for most protein pairs. Since there are fewer protein pairs remaining in the last three iterations and it becomes increasing difficult for the model to fit these data, feature importance changes drastically in latter rounds. By doing so, we can adjust the model parameters in each iteration and make new predictions based on the remaining protein pairs, and gradually finds the "outliers".

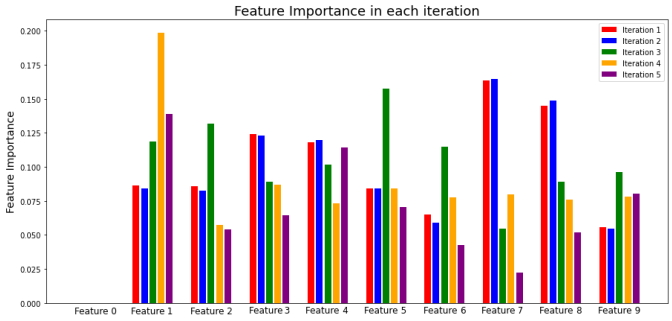


Fig. 8. Feature importance during each iteration for HEK293T dataset. It can be observed that the importance of the features changes greatly across iterations.

## V. Conclusion

We discover a brand-new application of CETSA data – protein-protein interaction (PPI) scores prediction. We use machine learning methods (e.g. Decision Tree) to predict the PPI scores for protein pairs in CETSA dataset and the predictions match the Bioplex ground truth very well. We further employ iterative clustering method and gradually reduce the number of mismatched protein pairs. At the end, only a small number of "outliers" (around $2\% - 3\%$ of total protein pairs) remains unfit, which are left for further biological investigation.

## References


[1] Martinez Molina D, Jafari R, Ignatushchenko M, Seki T, Larsson EA, Dan C, Sreekumar L, Cao Y, and P Nordlund, "Monitoring drug target engagement in cells and tissues using the cellular thermal shift assay," in *Science*, 2013, pp. 341(6141):84–7.

[2] Almqvist H, Axelsson H, Rozbeh J, Dan C, etc., and Mateus A, "Cetsa screening identifies known and novel thymidylate synthase inhibitors and slow intracellular activation of 5-fluorouracil," in *Nature Communications 7:11040*, 2016.

[3] Tsuyoshi Ishii, Takuro Okai, and Misa Iwatani-Yoshihara et al., "Cetsa quantitatively verifies in vivo target engagement of novel ripk1 inhibitors in various biospecimens," in *Nature Scientific Reports.*, 2017.

[4] Mayumi Kitagawa, Pei-Ju Liao, Kyung Hee Lee, Jasmine Wong, and See Cheng Shang et al, "Dual blockade of the lipid kinase pip4ks and mitotic pathways leads to cancer-selective lethality," in *Nature Communications.*, 2017.

[5] Miettinen TP, Peltier J, Härtlova A, and Gierliński M et al, "Thermal proteome profiling of breast cancer cells reveals proteasomal activation by cdk4/6 inhibitor palbociclib," in *EMBO J.*, 2018.

[6] Chernobrovkin AL, Legnqvist J, and et al, "In-depth characterization of staurosporin induced proteome thermal stability changes," in *bioRxiv*, 2020.

[7] Chris Soon Heng Tan, Ka Diam Go, Xavier Bisteau, Lingyun Dai, etc., and Pär Nordlund, "Thermal proximity coaggregation for system-wide profiling of protein complex dynamics in cells," in *Science.*, 2018 Mar 9, pp. TN.6380:1170–1177.

[8] Huttlin EL, Ting L, Bruckner RJ, Gebreab F, etc., and Gygi SP, "The bioplex network: A systematic exploration of the human interactome," in *Cell. 162 (2): 425–40.*, 2015.

[9] Huttlin EL, Bruckner RJ, Paulo JA, Cannon JR, etc., and Harper JW, "Architecture of the human interactome defines protein communities and disease networks," in *Nature. 545 (7655): 505–509.*, 2017.

[10] Banerjee S., Velásquez-Zapata V., Fuerst G., Elmore J.M., and Wise R.P., "a next-generation protein–protein interaction software," in *Briefings in Bioinformatics. 22 (4).*, Dec 2020.

[11] Wodak SJ, Vlasblom J, Turinsky AL, and Pu S, "Protein-protein interaction networks: the puzzling riches," in *Current Opinion in Structural Biology.*, December 2013, p. 23 (6): 941–53.

[12] LM Brettner and J Masel, "Protein stickiness, rather than number of functional protein-protein interactions, predicts expression noise and plasticity in yeast," in *BMC Systems Biology.*, September 2012, p. 6: 128.

[13] Ramachandran N, Raphael JV, Hainsworth E, and Demirkan G, "Next-generation high-density self-assembling functional protein arrays," in *Nature Methods.*, June 2008, p. 5 (6): 535–8.

[14] Bertolini M, Fenzl K, Kats I, Wruck F, Tippmann F, Schmitt J, and etc., "Interactions between nascent proteins translated by adjacent ribosomes drive homomer assembly," in *Science.*, Jan 2021, pp. 1;371(6524):57–64.

[15] Phizicky EM and Fields S, "Protein-protein interactions: methods for detection and analysis," in *Microbiological Reviews.*, March 1995, p. 59 (1): 94–123.

[16] Titeca Kevin, Lemmens Irma, Tavernier Jan, and Sven Eyckerman, "Discovering cellular protein-protein interactions: Technological strategies and opportunities," in *Mass Spectrometry Reviews.*, 29 June 2018, p. 38 (1): 79–111.