Protein prediction models support widespread post-transcriptional regulation of protein abundance by interacting partners

[1,2] Himangi Srivastava, **[1] Michael J. Lippincott, [1] Jordan Currie, [1] Robert Canfield, [1,2,3] Maggie P. Y. Lam, [1,2,]*Edward Lau**

[1]Department of Medicine/Cardiology
[2]Consortium for Fibrosis Research and Translation
[3]Department of Biochemistry and Molecular Genetics
University of Colorado School of Medicine
Aurora, CO 80045

***Correspondence:**
Edward Lau, PhD, edward.lau@cuanschutz.edu

Abstract

Protein and mRNA levels correlate only moderately. The availability of proteogenomics data sets with protein and transcript measurements from matching samples is providing new opportunities to assess the degree to which protein levels in a system can be predicted from mRNA information. Here we examined the contributions of input features in protein abundance prediction models. Using large proteogenomics data from 8 cancer types within the Clinical Proteomic Tumor Analysis Consortium (CPTAC) data set, we trained models to predict the abundance of over 13,000 proteins using matching transcriptome data from up to 958 tumor or normal adjacent tissue samples each, and compared predictive performances across algorithms, data set sizes, and input features. Over one-third of proteins (4,648) showed relatively poor predictability (elastic net $r \leq 0.3$) from their cognate transcripts. Moreover, we found widespread occurrences where the abundance of a protein is considerably less well explained by its own cognate transcript level than that of one or more trans locus transcripts. The incorporation of additional trans-locus transcript abundance data as input features increasingly improved the ability to predict sample protein abundance. Transcripts that contribute to non-cognate protein abundance primarily involve those encoding known or predicted interaction partners of

the protein of interest, including not only large multi-protein complexes as previously shown, but also small stable complexes in the proteome with only one or few stable interacting partners. Network analysis further shows a complex proteome-wide interdependency of protein abundance on the transcript levels of multiple interacting partners. The predictive model analysis here therefore supports that protein-protein interaction including in small protein complexes exert post-transcriptional influence on proteome compositions more broadly than previously recognized. Moreover, the results suggest mRNA and protein co-expression analysis may have utility for finding gene interactions and predicting expression changes in biological systems.

Introduction

Mounting evidence now shows that protein levels correlate imperfectly with the levels of their cognate transcripts [1–3]. More specifically, although a robust trend exists over the log scale between protein and mRNA measurements *across genes*, genewise correlation between proteins and their transcripts is much poorer *across observations* (samples, tissues, cell types, or subjects). This has been taken to indicate that while abundant proteins have abundant transcripts, transcript variance within a group of samples does not necessarily predict or signify corresponding protein changes [4]. Multiple factors are known to contribute to this non-correlation. Technical variations are often cited as a substantial source of non-correlation, as transcriptomics and proteomics measurements carry different sources of error and proteins with lower baseline variance in mass spectrometry have been shown to be better predicted by their transcripts [5]. Nevertheless, a substantial portion of protein variance remains unexplained and is likely attributable to biological and biophysical regulations. It has been well recognized that large multi-protein complexes could invoke a buffer effect on protein levels [6,7], as a multimeric complex only fully folds and functions when all subunits are present, any induction of the transcript for a single subunit would not per se lead to additional complexes, and resulting supernumerary proteins are thought to be quickly degraded [6,8]. Lastly, numerous post-transcriptional and post-translational mechanisms are known to modulate protein levels such as the gene- and context-dependent translation rates of mRNAs [9,10], the differential half-life and temporal distributions between mRNAs and proteins [2,11], and proteolytic degrading translated proteins in the cell.

The emergence of large-scale proteogenomics data from matching samples have created new opportunities to revisit protein-level predictions from transcriptomics data. The abundance of a protein may be the function of one or more transcripts. Most notably, available data sets from the Gene Tissue Expression (GTEx) project [7] and the Clinical Proteomic Tumor Analysis Consortium (CPTAC) [12], have spurred the use of machine learning approaches to evaluate how well one can predict protein level variance

given a set of transcriptomics data, with the goal of developing strategies and algorithms that can boost the performance of protein level predictions. This culminated in a community based effort in the CPTAC Proteogenomics Dream Challenge Task 2, which tasked participants with predicting protein abundances from mRNA and genetic data from CPTAC ovarian and breast cancer samples [12]. The results suggest that protein level prediction remains a challenging and not fully resolved problem, as many community submitted models did not improve substantially the baseline model, which is an elastic net taking into account all mRNA features available and has a median Pearson's correlation coefficient (*r*) of 0.47 for ovarian cancer. Nevertheless, general lessons have emerged from the top performing models; for instance: (i) ensemble methods generally performed well [12–14]; (ii) combining observations from the ovarian and breast cancer datasets to borrow information from each other led to improved predictions [12]; and (iii) judicious feature pre-selection based on prior biological knowledge such as protein-protein interactions improved prediction performance [12,14]. Notwithstanding these general observations, the current literature reflects that much remains to be learned about the relationship of mRNA and protein regulations in different genes and whether there are fundamental limits to how well mRNA abundance reflects that of their protein counterpart. This problem has several practical importances. Proteins carry out the majority of biological processes and hence are arguably the most relevant molecules to biological states. Despite rapid advances in proteomics techniques, bulk and single-cell RNA sequencing remain the most commonly used methods to interrogate gene expression status on a large scale and will likely remain so in the foreseeable future. Transcriptomics experiments often operate on the implicit assumption that identified differential regulation exert their biological effects via their cognate proteins, hence it is important to better understand the relationships between protein and mRNA levels to aid in data interpretation and determining potential protein level changes given a set of transcriptomics data. Alternatively, knowing the genewise difference in how well a gene's transcript can predict its protein counterpart may be useful for filtering and prioritizing biologically relevant transcript signatures [15].

Here we revisit the predictability of protein levels from transcriptomics data. Since the time of the Dream Challenge, considerably more proteogenomics data have been made publicly available which increases the number of observations available for modeling training, as well as the number of proteins for which there is mass spectrometry information available. Individual CPTAC cancer studies have analyzed the protein and mRNA correlation in individual tumors and normal adjacent tissues and nominated specific pathways whose correlations are particularly poor. Additional re-analysis and meta-analysis studies have outlined the distribution of prediction performances across algorithms, and generally conclude there is some statistical

enrichment of biological processes or protein features among proteins that are poorly predicted by their own transcript level, e.g., metabolic and essential proteins or proteins belonging to complexes [12,16]. Nevertheless, a granular analysis remains unrealized in the literature that interrogates the identity and regulatory modality of individual proteins in depth. Accordingly, our goals here are to (1) evaluate how the increasing data size from combining CPTAC tumor data sets affects the performance of prediction algorithms and feature selection strategies; and (2) interpret prediction models to assess the importance of transcript features in individual protein abundance regulation. The results suggest that the incorporation of transcript level information from protein interacting partners played a substantial role in predicting protein levels, and moreover, there are widespread instances in the proteome where the abundance of a protein correlates primarily with a trans locus transcript than its own cognate transcript, which has implications for gene expression profiling studies.

Methods

Data retrieval and processing

Gene expression data were obtained from public data from the CPTAC project and included data from 8 cancer types: ovarian cancer (OV) [17], breast cancer (BR) [18], endometrial carcinoma (EN) [19], colorectal cancer (CO) [20], lung adenocarcinoma (LUAD) [21], clear cell renal carcinoma (CCRCC) [22], glioblastoma (GB) [23], and lung squamous cell carcinoma (LSCC) [24]. The cumulative inclusions of each cancer type in the order above are sequentially referred to as CPTAC_2 to CPTAC_8 in the manuscript, such that CPTAC_2 refers to the union of ovarian and breast cancer (OV + BR); CPTAC_3 refers the union of ovarian, breast, and endometrial cancer (OV + BR + EN); and so on. The mRNA and protein level expression data from the CPTAC cancer types was retrieved using the cptac package v.0.9.7 [25] in Python 3.9. Each column of the quantitative measurement of the transcriptomics data acted as an independent variable or feature variable whereas the normalized quantitative measurement of a particular protein of interest acted as the single dependent or target variable in the protein model. Retrieved mRNA level gene expression data are standardized using the scikit-learn simple scaler. The proteomics data were likewise downloaded using the cptac package as presented in the data, and were stable isotope labeled relative quantitative mass spectrometry data presented as normalized log ratios across samples as in the original studies. All tumor samples were labeled using Thermo tandem mass tag (TMT) 10- or 11- plex isobaric tags for MS2 quantification, with the exception of the ovarian cancer data, which were labeled with Sciex iTRAQ isobaric

tags for MS2 quantification, and the colon cancer data, which contained both label-free and TMT quantifications. The retrieved log ratios across samples were not further transformed.

For each protein for which predictions are to be made, we retrieved five separate feature sets:

1. **Single:** Using only the single transcript coding for the protein of interest for model training, then running the pipeline to train a model for each protein separately.
2. **CORUM:** Using the transcripts of all proteins belonging to the same protein complex as the protein of interest, if any, in CORUM v.3.0 [26], where protein complexes are defined as two or more proteins that interact physically in a quaternary structure. These transcripts then act as independent variables (features) to predict the target variable (protein of interest). The pipeline is then run to train a model for each protein separately.
3. **STRING 800:** Using the transcripts of interacting partners of the protein of interest as input features. Interacting partners are retrieved from STRING v.11 [27], which documents functional associations including physical interactions, genetic interactions, co-expression, co-occurrence, and other associations. The STRING combined score represents the overall likelihood of interactions. Interacting pairs with a STRING combined score of 800 or above (high-confidence) are included. The pipeline is then run to train a model for each protein separately.
4. **STRING 200:** As above, except that interacting pairs with a STRING combined score of 200 or above (low- to high-confidence) are included. The pipeline is then run to train a model for each protein separately.
5. **Transcriptome:** A transcriptome-wide model where all qualifying transcripts in the data set are included as features, prior to the removal of low-variance features. The pipeline is then run to train a model for each protein separately.

Model training and evaluation

For each feature set, missing values for feature variables are imputed using median imputation, followed by the removal of features with variance of 0.2 or below. Models are not trained for proteins with fewer than 50 empirical observations. The data are then split 80:20 into training and test sets. No imputation or additional standardization was performed on the proteomics data. The input and target data from the training set are then used to train a model using either linear regression, elastic net with 5-fold cross validation, random forest regressor, or gradient boosting regressor in scikit-learn v.1.0 [28], with the following specified parameters: random forest number of estimators: 500, criterion: squared error, max depth: 4; elastic net cross validation L1 ratio: 0.1, 0.5, 0.9, 0.95, cv: 5, tolerance: 1e–3, max iterations: 2000; gradient boosting:

n_estimators: 1000, max_depth: 3, subsample: 0.5, min_samples_splot: 5, learning_rate: 0.025. The trained models are saved as individual objects which include the predicted protein levels in the training set as well as the contribution of each feature to the overall prediction (coefficients in linear regression and elastic net; feature importance and trees in random forest and gradient boosting regressors), and are applied to predict protein levels in the test set data. Further interpretation of feature importance was performed using Shapley values with the aid of the shap package v.0.40.0 [29], or with the Boruta algorithm using the Boruta_Py package v.0.3 [30].

To evaluate model performance, the Pearson's correlation coefficients between predicted ($\hat{y}$) and actual (y) protein values are calculated individually for the train and test set data for each protein model using the numpy corrcoef function, or defaulted to 0 if the standard deviation of predicted values is 0. Goodness-of-fit (R^2) is calculated using scikit-learn.metrics.r2_score function. Normalized root mean square errors (NRMSE) are calculated as follows:

$$NRMSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1} (y_i - \hat{y}_i)^2 * \frac{1}{(max(y) - min(y))}}$$

Network construction and analysis

To construct the protein regulation network, we first constructed an overall graph from a subset of transcript-protein relationships using the networkx package [31] in Python 3.9. Qualifying proteins are those whose prediction increased with the inclusion of more features, such that in the STRING or CORUM feature set, the correlation coefficient between the elastic net predicted and actual protein levels is greater than the single feature elastic nets by at least 0.25. Transcripts that contribute to the prediction of these proteins are therefore included if their elastic net coefficient *a* and random forest feature importance *b* are greater than a certain threshold, which was set to *a*=0.05, *b*=0.05 for the CORUM feature set and *a*=0.2, *b*=0.05 for the STRING feature set.

In the data, the list of proteins and their corresponding transcriptomics features act as the nodes and the interaction pattern is a function of the correlation value and feature importance between a protein and the transcriptomics. The direction of the edge therefore flows from transcriptomics to protein. All the text files for each protein obtained from the computational pipeline are compiled into a n overall data frame. This data

frame is then converted into the form of a network data frame. In this data frame, all the protein becomes the target node, and the transcriptomics corresponding to the protein becomes the source node. Additional column of this data frame consists of the weight between a protein and its predicting transcripts, which was calculated using the sum of the [–1,1] clipped elastic net coefficient and the descending percentile rank of the random forest feature importance. An overall directed graph G(V, E) is then constructed from the edge lists with proteins as targets and their predicting transcripts as sources, such that it is a function of vertices V which depict protein and/or transcript nodes and edges E which flow in the direction from the transcript sources to the protein targets. Using the built-in functions in networkx, this overall directed graph is divided into weakly connected components, which resulted in individual subgraphs such that for the overall network G=(V, E) where V is the vertices and E is the edge then a subgraph of G = (V, E) is a graph S=(V′, E′) where vertex set $V' \subseteq V$ and edge set $E' \subseteq E$ connects only nodes of V′. We then used Cytoscape v.3.9 [32] to visualize and perform additional topology analysis and functional annotations of each subgraph. HiDef persistent community detection [33] was performed with the aid of the CyCommunityDetection Cytoscape plugin [34] using the Leiden algorithm [35]. Functional enrichment of individual communities was performed using CyCommunityDetection with the g:Profiler web server [36]. Hub nodes were identified with the aid of the CytoNCA [37] Cytoscape plugin using the betweenness centrality algorithm with the top 10% of proteins with the highest centrality defined as hubs. Functional enrichment analysis on the hub proteins was performed with the aid of the stringApp [38] Cytoscape plugin using default settings and 5% false discovery rate cutoff.

Additional data analysis

Additional data analysis, statistics, and visualization were performed in R v.4.1.1 with the aid of the ReactomePA [39], clusterProfiler [40], and circlize [41] packages; and in Python 3.9 with the aid of the seaborn [42] package.

Data and code availability

Code to train machine learning models and evaluate prediction performance have been uploaded to GitHub at https://github.com/Lau-Lab/CPTAC_Protein. Code for making figures can be found under notebook/PLOS_MakeFig*.qmd.

Results

Feature selection improves the prediction of protein abundance from transcriptome data

We first evaluated whether increasing proteogenomics data depth would impact optimal feature selection strategies and algorithms in prediction protein levels. To do so, we retrieved data from up to 8 tumor types from the CPTAC data set and processed the data to collage a data table with matching protein and transcript data from each tumor sample . The data sets were ordered such that breast and ovarian cancers as in the DREAM challenge were combined first, then other cancer types were added according to the order of their availability (see Methods). We trained models to predict the target variable (the normalized labeled mass spectrometry measured abundance of a particular protein) from various input features (transcriptomics data from the matching samples) using either a multiple linear regression, elastic net, and random forest regressor in scikit-learn. To compare the performance of prior knowledge based feature selection, we further compared supplying the models with input features based on (1) only RNA level data the corresponding transcript of the target variable protein (single feature) (2) self transcript plus the transcript of any protein within the same protein complex of the target variable protein in CORUM (CORUM feature); (3) self transcript plus the transcript of any proteins that are high-confidence protein interaction partners including physical interactions and other inferred association from STRINGdb (STRING 800 feature); (4) self transcript plus the transcript of any proteins that are low to high- confidence protein interaction partners from STRINGdb; (5) using all qualifying transcripts (transcriptome features).

We next compared the genewise prediction performance of the transcriptome using the test set correlation coefficient between predicted and actual mass spectrometry protein values, R^2, and NRMSE as described (**Figure 1A; Supplementary Table S1**). The elastic nets and random fores models performed better with more features and more data. In contrast, a multiple linear regression model performed little better with more included features and in fact failed to predict protein levels when a large number of features are given. For the elastic net and random forest models, performance gains began to saturate with additional data sets (r: 0.388 and 0.402, respectively) but further increased when larger feature sets were given (r: 0.599 for both) (**Figure 1A; Supplementary Figure S1**). As a comparison, we also evaluated the performance of the models in single cancer data (i.e., in each data set alone) (**Supplementary Figure S2A**) as well as combined the data sets in the order of descending performance (**Supplementary Figure S2B**). The single cancer models suggest that a prediction performance of up to median r of 0.691 is achievable using the Transcriptome wide feature set, and 0.69 in the STRING 200 feature set (random forest, CCRCC),

albeit in a single cancer type only. In both the single cancer type and reordered combination comparisons, the observation remains true in all cases that the inclusion of protein interactor features further increased the prediction performance of both the elastic net and random forest algorithms. Hence, substantial prediction improvements resulted from the inclusion of other transcripts (number of features) over the gain in data set sizes (number of observations) or the use of algorithms (random forests) that can account for non-linearity. This is corroborated when considering the feature sets across the largest data set collection used (CPTAC_8 from 8 cancer types). Model performance continued to increase as more features were added in (median r from 0.398 to 0.599), and the number of non-predictors and negative predictors decreased while overall dispersion of predictability also decreased.

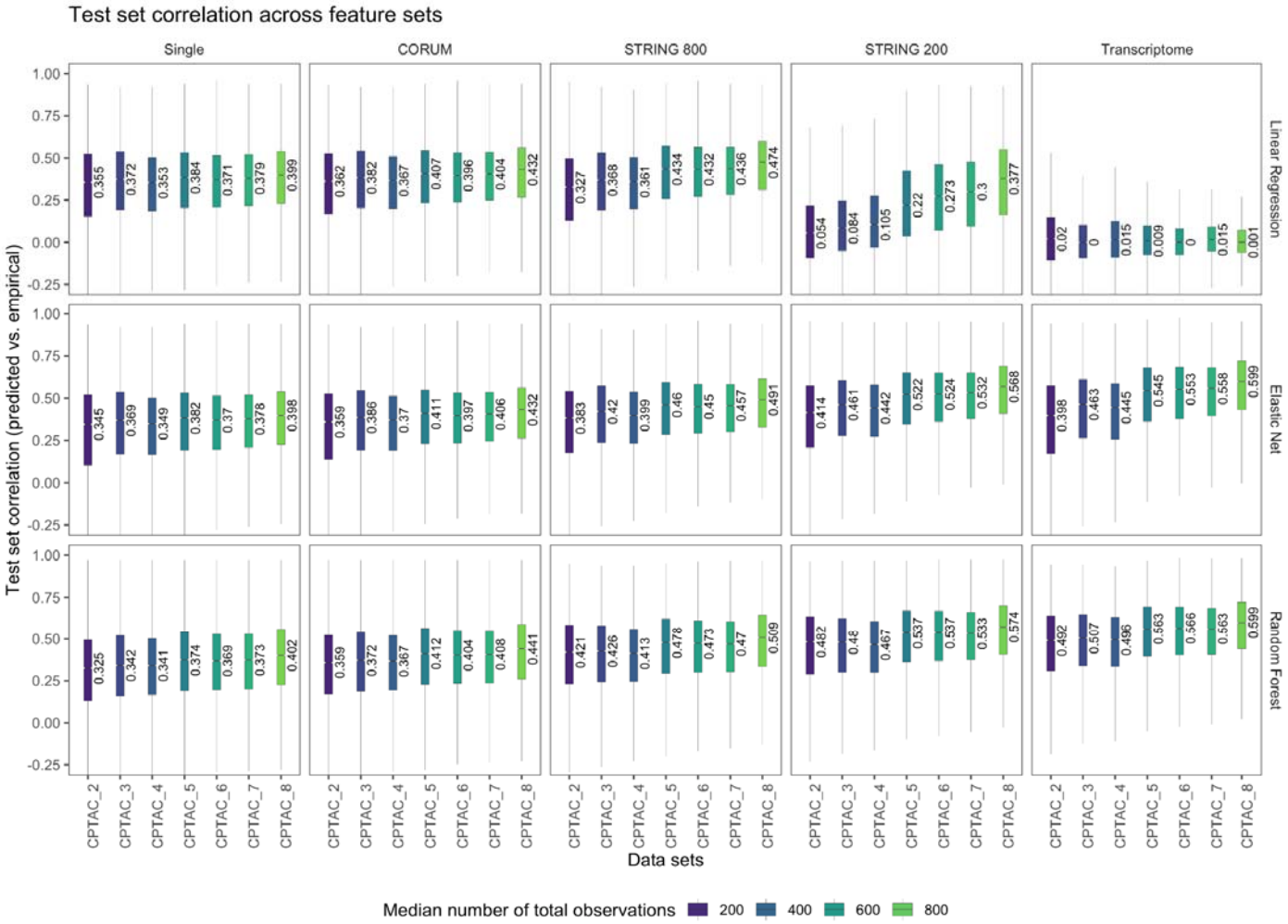


Figure 1. Genewise dispersion of protein predictability from transcriptome data. Box plots of test set correlation coefficients between the transcript-predicted and actual protein level for each protein are shown across five feature sets (column: single/self transcript, CORUM interactors, STRING 800 high-confidence associated proteins; STRING 200 low-confidence associated proteins, and all transcripts) and three algorithms (multiple linear regression, elastic net, and random forest). In each plot, the x axis denotes the number of additive CPTAC data sets used to train the models as described in Methods; box: interquartile range; whiskers: +/– 1.5 IQR; notch: SEM.

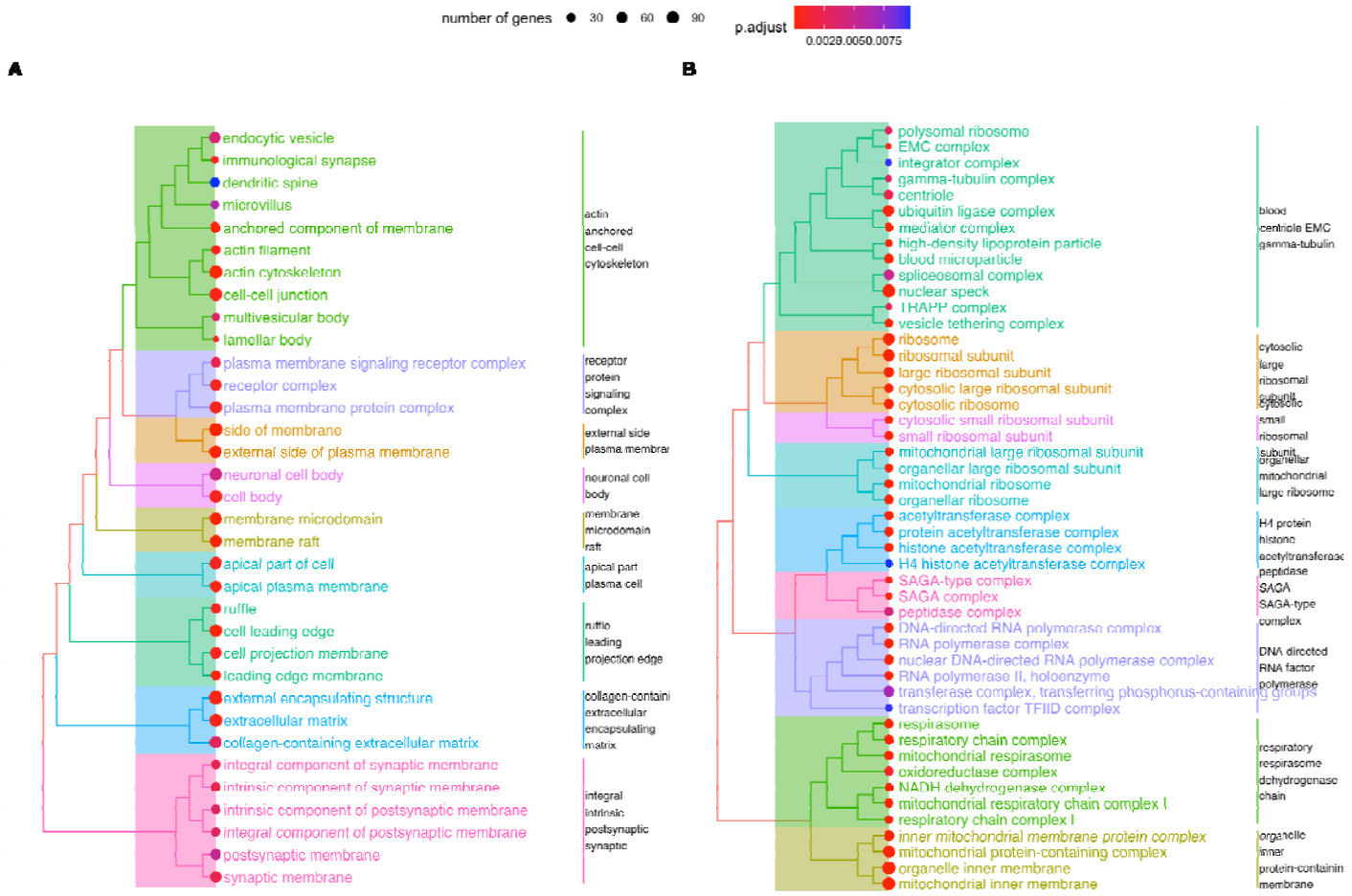


Figure 2. Pathway enrichment of proteins with good and poor predictability. A. Tree plots showing the clustering and relationships of gene ontology terms that are significantly enriched among proteins whose abundances are well predicted by their own transcripts (r ≥ 0.6). **B.** Tree plots of terms enriched among proteins whose abundances are poorly predicted by their own transcripts (r ≤ 0.3).

From the baseline self-transcript model in the CPTAC_8 data set containing 8 cancer types, we found that strong protein predictors (r ≥ 0.6; 2,008 genes) are enriched in genes participating in membrane proteins and cell junctions (**Figure 2A; Supplementary Table S2**), whereas poor protein predictors (r ≤ 0.3; 4,648 genes) are enriched in multiple large multi-protein complexes (**Figure 2B; Supplementary Table S3**). Hence the analysis of CPTAC_8 data set confirms prior observations that protein complex membership presents a major source for transcript-protein non-correlation.

We next compared the genewise performances between the STRING feature elastic net model and the self transcript model to investigate the observed performance gain at a more granular level. Considering the 13,239 genes with protein-level prediction and with at least one STRING interactor, on average, each gene in the STRING model has a median of 273 [158–457] features, compared to 1 feature in the self transcript model. Incorporation of STRING features led to an average increase in test set correlation coefficients of 0.16 [IQR: 0.07–0.27] (**Supplementary Table S4**). There is a significant positive correlation between the number of features and prediction improvements (Pearson's r 0.21, P: 8.2e–131) (**Supplemental Figure S3A**), a relationship which persisted when only genes with 10 or more interactors were included (**Supplemental Figure S3B)**. A similar observation is seen in the CORUM feature set where the number of transcript features (i.e., CORUM complex interactors) of a particular protein is positively correlated with improvement in predictability of protein abundance (**Supplemental Figure S3C–D)**. A functional enrichment analysis of proteins with strong improvements in prediction (difference in STRING vs. self-feature in test set predicted-actual correlation coefficient (Δr) ≥ 0.25) showed a strong enrichment in large multi-protein complexes including the ribosome, mitochondrial ribosome, RNA polymerase, and spliceosome (**Supplementary Table S5**) similar to the results from poor predictors. This corroborates that the non-correlation between transcripts and proteins of multi-protein complexes levels can be partially rescued by considering the information of interacting partner transcripts, which could participate in post-transcriptional regulations such as by being the stoichiometrically limiting transcript or crucial assembly parts, such that the transcript level of one subunit can regulate the protein level of other subunits.

Overall, these results are consistent with information about protein level residing both in the transcripts of the genes encoding the protein as well as the protein association partners. Moreover, taking into consideration the number of features in the STRING 800 and STRING 200 sets (median feature sizes of 10 [IQR: 2–48] and 266 [IQR: 149–451], respectively) compared to the size of the whole transcriptome feature set, the results indicate that a substantial portion of performance gains over the single-transcript feature set is already achievable from relatively few selected prior features.

Proteins whose abundances are predicted by non-cognate transcripts are common

We next examined more closely the underlying causes behind the performance gains of protein predictions in the CORUM and STRING feature sets. To do so, we examined the genes whose protein prediction performance from transcripts increased substantially (Δr ≥ 0.25) after the incorporation of additional transcript features. In total, we observed 484 such proteins in the CORUM feature sets, and 3,272 proteins from the STRING data set, representing over 24% of all examined proteins. These numbers increase further when a less conservative Δr threshold of ≥ 0.15 is used (6,123 proteins in the STRING dataset, 946 in CORUM), altogether suggesting non-cognate transcript contributions to protein level are common at the proteome level. Upon inspecting the model feature coefficients (elastic nets) and feature importance (random forests), we observed that a considerable portion of these proteins with improved prediction are associated with (1) poor contribution from the self-transcript, and (2) a substantial contribution from primarily a few non-cognate (i.e., trans locus) transcripts coding for other proteins (see below). In other words, although model performance continued to increase with increased feature set sizes, the contributions of trans locus transcript features to overall prediction performance is unevenly distributed, and are therefore attributable to a few high contribution genes rather than a simple scaling with feature size. Notably, despite the STRING feature set being substantially larger than the CORUM feature set, substantial contributions from trans locus transcripts in the STRING feature set primarily involve transcripts encoding proteins that form part of the same CORUM complex as the protein of interest itself, suggesting that stable complex memberships play an outsized role in determining protein levels.

To illustrate the poor predictive power of cognate transcripts on some proteins, we next interrogated a subset of proteins where a non-self transcript has an outsized effect on the protein level of the proteins in the CORUM and STRING feature sets: Propionyl-CoA Carboxylase Subunit Beta (PCCB), C-X9-C Motif Containing 1 (CMC1), Proteasome Assembly Chaperone 2 (PSMG2), SMCR8-C9orf72 Complex Subunit (SMCR8), Mitochondrial Calcium Uptake 2 (MICU2), and Protein Phosphatase 3 Regulatory Subunit B, Alpha (PPP3R1). PCCB forms the propionyl-CoA carboxylase complex with two protein members, which breaks down certain amino acids in the cell (**Figure 3A**). PCCB showed poor prediction by their own self-transcripts (r:–0.008) but experienced high gains in test set correlation upon incorporating additional transcript features in the CORUM (r: 0.483) and STRING (r: 0.692) models. CMC1 is an assembly factor that

forms an early intermediate of the cytochrome c oxidase complex in the mitochondrion with 14 documented protein members in CORUM (**Figure 3B**). Likewise, CMC1 showed only moderate prediction by their own self-transcripts (r: 0.274, respectively) but experienced high gains in test set correlation upon incorporating additional transcript features in the CORUM (r: 0.615) and STRING (r: 0.564) models.

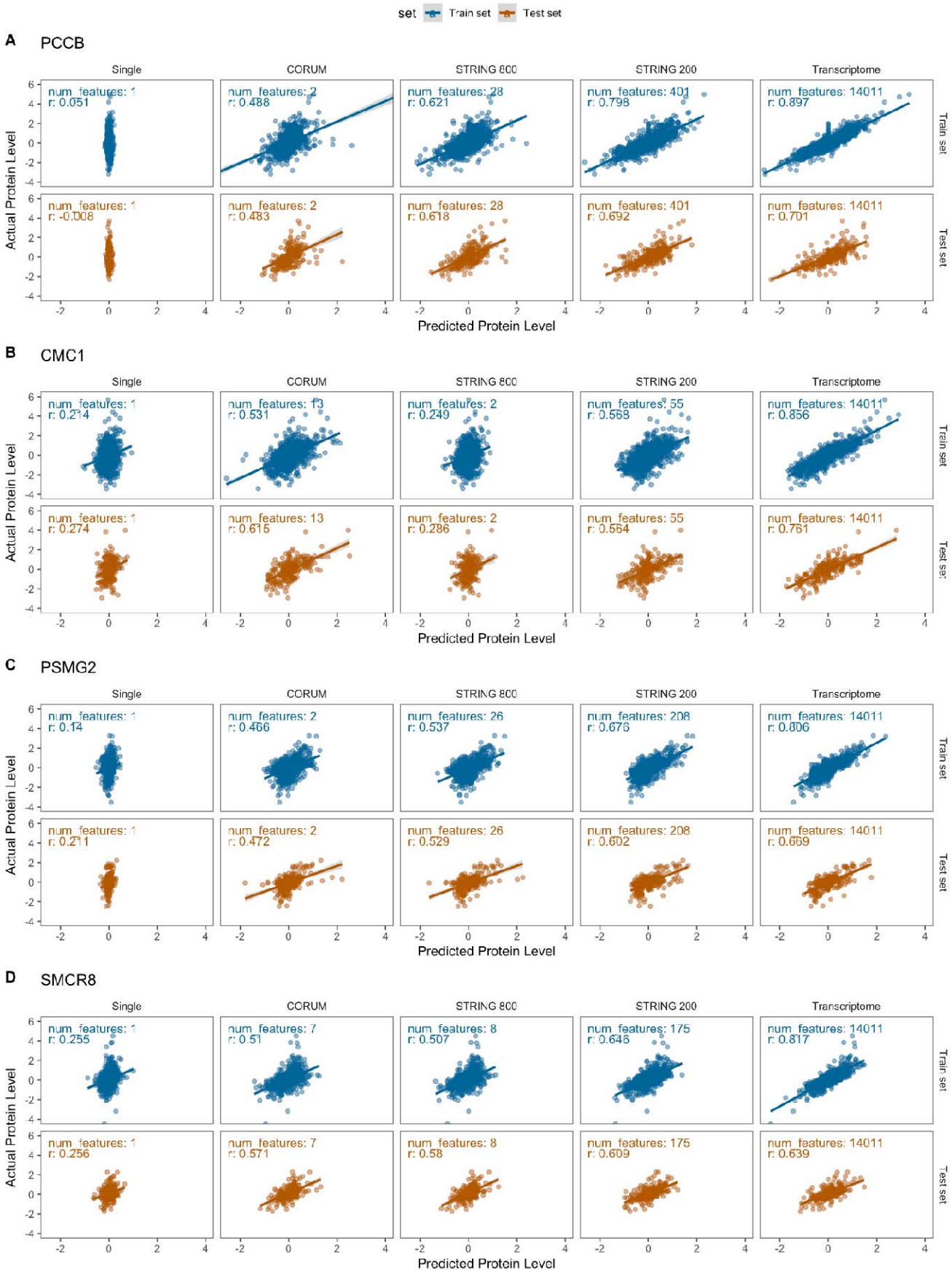

set Train set Test set
A PCCB
Single
CORUM
STRING 800
STRING 200
Transcriptome
num_features: 1 r: 0.051
num_features: 2 r: 0.488
num_features: 28 r: 0.621
num_features: 401 r: 0.798
num_features: 14011 r: 0.897
num_features: 1 r: -0.008
num_features: 2 r: 0.483
num_features: 28 r: 0.618
num_features: 401 r: 0.692
num_features: 14011 r: 0.701
Train set
Test set
Actual Protein Level
Predicted Protein Level
B CMC1
num_features: 1 r: 0.214
num_features: 13 r: 0.531
num_features: 2 r: 0.249
num_features: 55 r: 0.568
num_features: 14011 r: 0.856
num_features: 1 r: 0.274
num_features: 13 r: 0.615
num_features: 2 r: 0.286
num_features: 55 r: 0.564
num_features: 14011 r: 0.761
C PSMG2
num_features: 1 r: 0.14
num_features: 2 r: 0.466
num_features: 26 r: 0.537
num_features: 208 r: 0.676
num_features: 14011 r: 0.806
num_features: 1 r: 0.211
num_features: 2 r: 0.472
num_features: 26 r: 0.529
num_features: 208 r: 0.602
num_features: 14011 r: 0.669
D SMCR8
num_features: 1 r: 0.255
num_features: 7 r: 0.51
num_features: 8 r: 0.507
num_features: 175 r: 0.646
num_features: 14011 r: 0.817
num_features: 1 r: 0.256
num_features: 7 r: 0.571
num_features: 8 r: 0.58
num_features: 175 r: 0.609
num_features: 14011 r: 0.639

Figure 3. Proteins with improved predicted levels after inclusion of additional transcript features. Four proteins with substantial predictability from transcriptome data upon the inclusion of additional features are shown: **A.** PCCB, **B.** CMC1, **C.** PSMG2, **D.** SMCR8. For each protein, the transcript-trained prediction of protein level is plotted on the x axis and the actual protein level is plotted on the y axis. The lack of variance in predicted protein levels from the self-transcript model is due to the regularization of the elastic net model, and corresponds to a lack of correlation between PCCB mRNA and protein (see Figure 4). Blue: train set, brown: test set. Columns denote the transcript feature set used to train the model. The number of features used to train the model in each feature set is shown inside each plot. r: Correlation coefficient.

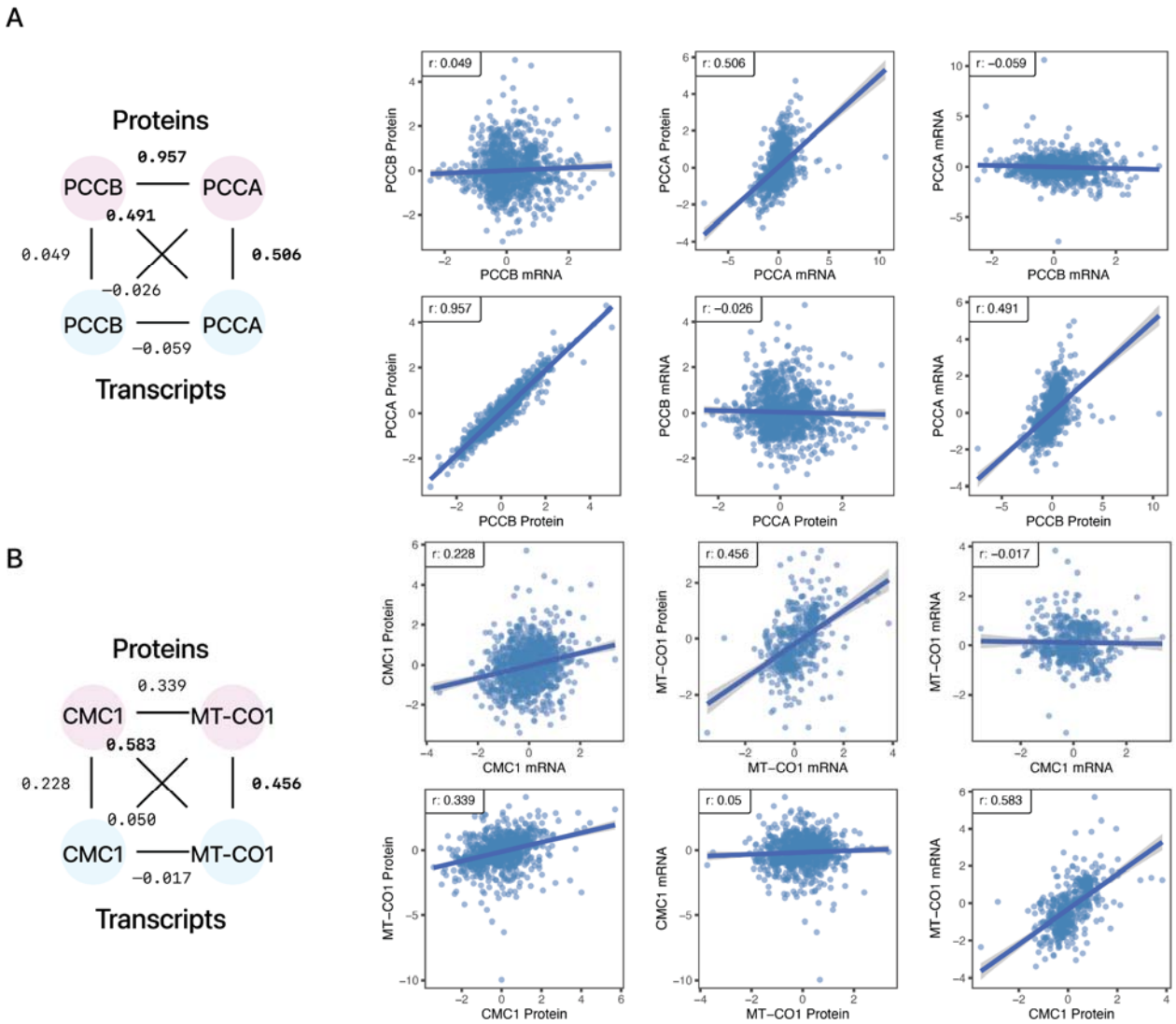


Figure 4: mRNA-Protein correlations of PCCB and CMC1 with functionally associated proteins. Two examples of proteins whose abundance is better explained by another transcript are shown. **A.** PCCB protein level is predicted by PCCA transcript but not its own transcript. **B.** CMC1 protein level is explained by MT-CO1 transcript level but not its own transcript. Substantial correlations across transcripts and proteins (≥ 0.4) are bolded.

We then considered the transcriptomics and proteomics data distributions of the features that had highest coefficients in the elastic net model from the CORUM feature set for PCCB and CMC1. For PCCB, most protein-level predictions are recovered when the feature set includes PCCA, which together with PCCB forms the stable propionyl-CoA carboxylase enzyme that consists of six copies of PCCB and six copies of PCCA each. There is a remarkably strong correlation with its propionyl-CoA carboxylase complex interacting partner PCCA at the protein level (r: 0.957), but this correlation is almost entirely absent at the transcript level (r: –0.059). This is partially explainable by the observation that PCCA transcript has a higher level of variability and is also strongly correlated with the PCCB protein level. The total protein level of PCCB is therefore primarily driven by PCCA rather than PCCB transcripts (**Figure 4A**).

In the case of CMC1, we found that MT-CO1 has the highest contribution to CMC1 protein level (**Figure 4B**). MT-CO1 is the mitochondrial genome encoded subunit of complex IV that is thought to act as the nascent scaffold around which the complex is assembled. CMC1 binds to MT-CO1 during the formation of the early module known as MITRAC and is subsequently released during assembly. [43,44] CMC1 transcript and protein levels are moderately correlated with one another (r: 0.228). Again, CMC1 and MT-CO1 show poor co-expression at the transcript (r: –0.017) level and an improved correlation at the protein (r: 0.339) level, whereas MT-CO1 transcript is robustly correlated with the CMC1 protein level (r: 0.583), which suggests the possibility that proteogenomic co-expression may reveal additional functionally related protein pairs. Closer inspection suggests the possibility of a non-linear relationship between MT-CO1 transcript and CMC1 protein where upon reaching a plateau, further increases in MT-CO1 transcript does not correlate to further increase in CMC1 protein levels, suggesting other transcripts may also contribute to CMC1 protein levels. Consistent with this, other complex IV subunits also have non-negligible coefficients (in the elastic nets) or feature importance (in the random forests) in the CMC1 model. When the feature set expanded to include all STRING proteins, MT-CO1 remained a high contributor in the random forest but not the elastic net model, suggesting the elastic nets may be more susceptible to collinear features than the random forests.

Two other examples of small stable complexes with interdependent protein and mRNA correlation are shown in **Figure 5.** Proteasome Assembly Chaperone 2 (PSMG2) forms the heterodimer PAC1-PAC2 complex (2 protein members in CORUM) along with Proteasome Assembly Chaperone 1 (PSMG1), which binds with 20S proteasome precursors and acts as a scaffold to promote proteasome ring assembly while preventing aberrant dimerization of 20S proteasome α rings [45]. PSMG2 and PSMG1 are co-expressed strongly at the protein level (r: 0.713) but only moderately at the transcript level (0.222) (**Figure 5A)**. PSMG2 protein has a higher correlation with PSMG1 transcript (r: 0.472) than its own (r: 0.158). From GTEx v8 data, PSMG2 is expressed more highly than PSMG1 across tissues (TPM ~40 vs. 15) [46], which suggests PSMG1 may act as a limiting factor in heterodimer formation. PSMG1 knockout in mice has been shown to decrease PSMG2 at the protein level, consistent with our finding that PSMG1 transcript levels influence PSMG2 protein abundance [47].

In another example, Smith-Magenis Syndrome Chromosomal Region Candidate Gene 8 (SMCR8) forms the heterotrimer C9orf72-SMCR8-WDR41 complex (3 protein members in CORUM) together with the C9orf72-SMCR8 Complex Subunit **(**C9orf72) and WD Repeat Domain 41 (WDR41) [48]. C9orf72-SMCR8 is known to dimerize prior to binding with WDR41 to form the heterotrimer. The C9orf72-SMCR8 complex modulates autophagy, likely by modulating the maturation of autophagosomes [49]. Mutations in the genes encoding the C9orf72-SMCR8 Complex are broadly implicated in diseases including amyotrophic lateral sclerosis and frontotemporal dementia. As in the other cases, we observed a co-expression of SMCR8 and C9orf72 at the protein level (0.632) but not the transcript level (–0.012), and SMCR8 protein levels correlate more strongly with the C9orf72 transcript (r: 0.436) than the SMCR8 transcript (r:0.231) (**Figure 5B**). A previous study suggested C9orf72 may stabilize excess SMCR8 [50], providing one possible explanation for the mechanism behind the observations here.

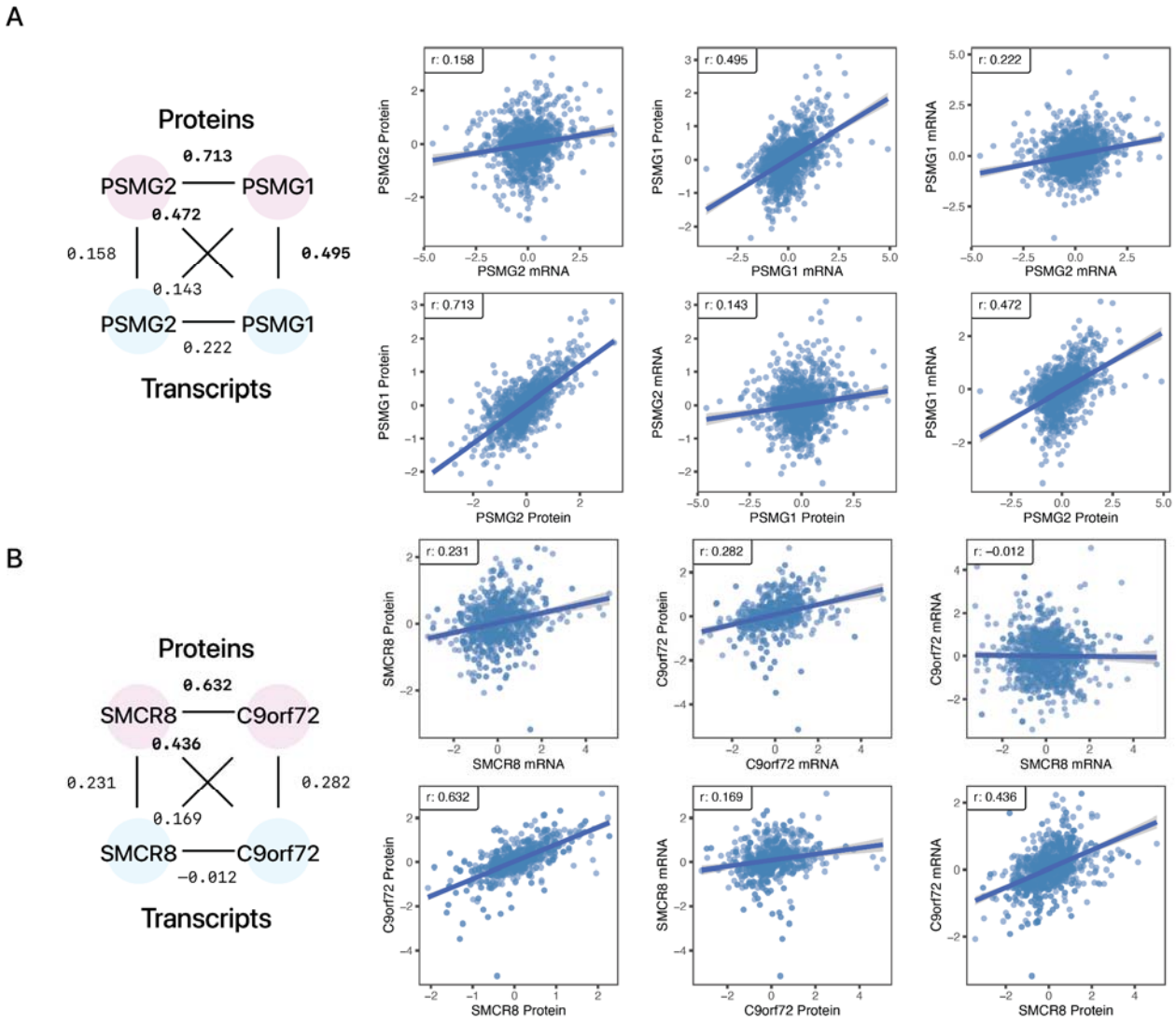


Figure 5: mRNA-Protein correlations of PSMG2 and SMCR8 with functionally associated proteins. Two examples of proteins whose abundance is better explained by another transcript are shown. **A.** PSMG2 protein level is predicted by PSMG1 transcript but not its own transcript. **B.** SMCR8 protein level is explained by C9orf72 transcript level but not its own transcript. Substantial correlations across transcripts and proteins (≥ 0.4) are bolded.

Supplementary Figure S5 shows two more examples concerning small complexes. The mitochondrial calcium uniporter complex (MCU complex) contains five protein members in CORUM (**Supplementary Figure S5A**). Regulatory subunits of the complex, including calcium uptake protein 2,

mitochondrial (MICU2) sense calcium levels, regulating the uptake of calcium performed by the mitochondrial calcium uniporter (MCU) [51]. Heterologous overexpression of FLAG-tagged MCU has been shown to increase expression of MICU2 in culture [52]. This is interesting, given the observation here that at an endogenous level of MCU expression, MICU2 protein is more strongly correlated to the MCU transcript (r: 0.488) than its own (r: 0.058) (**Supplementary Figure S5A**).

The final example highlighted here comes from the heterodimer calcineurin, which is made up of a catalytic subunit (calcineurin subunit A), with 3 isoforms encoded by 3 separate genes (PPP3CA, PPP3CB and PPP3CC), [CJ2] and a Ca^{2+} binding regulatory subunit (calcineurin subunit B), with 2 isoforms also encoded by 2 separate genes (PPP3R1 and PPP3R2) [53]. Calcineurin is a serine/threonine phosphatase which modulates many calcium dependent signaling pathways, and is the target of immunosuppressant drugs cyclosporin A and FK506 [54]. In the analysis here, calcineurin subunit B type 1 (PPP3R1) protein correlated poorly with PPP3R1 transcript levels (r: 0.205), but strongly correlates (r: 0.819) with the protein phosphatase 3 catalytic subunit alpha (PPP3CA) transcript — an isoform of calcineurin subunit A (**Supplementary Figure S5B**). Together, these cases illustrate a general observation — in cases where the protein abundance of a gene is more highly correlated to the transcript level of a binding partner than to its cognate transcript, there appears to be stronger co-expression between the gene and its abundance driver at the protein level but not at the transcript level.

To corroborate these observations, we used the Boruta algorithm against the random forest CORUM models and confirmed each of the principal non-cognate transcript features is retained. We further considered Shapley values as an explanation of random forest and gradient boosting models, and likewise found that the interacting protein transcripts formed the top contributors in explaining the abundance of the protein of interest in each of the six cases above (**Supplementary Figure S6**). In some examined cases including PSMG2, the transcript coding for the protein of action is more abundant than that of its regulating interacting partner, which is compatible with a scenario where simple stoichiometric constraints control the protein level of the supernumerary subunit. It is however less clear whether this scenario applies to other examined cases including CMC1 and MT-CO1, where the MT-CO1 transcript is much more abundant than CMC1 owing to the multiple copies of mitochondrial genome in the tissue. Hence, how MT-CO1 regulates CMC1 protein level is not explainable directly from transcript stoichiometry alone and awaits further mechanistic studies. Finally, we also examined feature importance in a single data set of a recent CPTAC study that showed the highest performance in the single feature model (LSCC) (**Supplementary Figure S7**).

Among the six highlighted proteins (PCCB, CMC1, PSMG2, SMCR8, MICU2, PPP3R1), five were best predicted by a non-self transcript, and five of the top trans locus predictors were conserved from the combined data set analysis with the exception of MT-CO1, which was not among the model features due to the number of shared observations required. Hence the observed trans locus predictors are conserved in a single data set model and are unlikely to be due to heterogeneity across data sets.

Table 1 summarizes the top 50 genes whose protein abundance becomes substantially more predictable upon the inclusion of CORUM features, along with their top protein level contributor transcripts. Notably, a number of these genes are found in multiple gene sets within the MSigDB C2 CGP gene set collection that are commonly used for functional enrichment analysis [55], whereas others are found to be significantly associated with diseases in the literature from bibliometric analysis [56,57], hence these genes are common and important to multiple areas of biomedical inquiries. We suggest the possibility that the transcripts of these genes may not present accurate proxy variables for their proteins should be considered in linking transcript level changes to downstream cellular physiology.

Taken together, these results exemplify widespread interdependency of protein levels on trans locus transcripts, involving not only large megadalton protein complexes but many small complexes involved in diverse biological processes in the cell. The CORUM feature set considered here alone contains over 2,700 human complexes derived from CORUM release 3.0 mappable to 3,689 proteins with distinct gene names, with a median of 3 proteins per complex (**Supplementary Figure S8**). Thus, small protein complexes are widespread in the proteome and a large fraction of the proteome could be placed under complex post-transcriptional control that decouples protein levels from mRNA levels. Moreover, half of the proteins in the CORUM annotations belong to 2 or more distinct complexes (including intermediates and subcomplexes) which suggest further opportunities for more complex patterns of interdependent protein and mRNA levels.

Gene	Gene name	Representative CORUM complex name	Test set r, single feature	Test set r, CORUM features	# CORUM feature	Top non-self contributor	# MSigDB CGP sets	Most significant Disease Ontology association (P < 0.05)
APOL1	apolipoprotein L1	APOL1 complex B (APOL1, APOA1, HPR, FN1, IGHM)	0.118	0.435	5	APOA1	20	Kidney disease (P=0.00011)
BORCS8	BLOC-1 related complex subunit 8	BORC complex	0.189	0.456	8	BORCS7	5	
CACNA1A	calcium voltage-gated channel subunit alpha1 A	G protein complex (CACNA1A, GNB1, GNG2)	0	0.305	3	GNG2	31	Familial hemiplegic migrane (P=0.00032)
CAPZB	capping actin protein of muscle Z-line subunit beta	CAPZA3-CAPZB complex	0.032	0.719	22	WASHC2A	19	
CMC1	C-X9-C motif containing 1	Cytochrome c oxidase, mitochondrial	0.274	0.615	13	MT-CO1	10	Osteoarthritis (P=0.002)
COL18A1	collagen type XVIII alpha 1 chain	ITGA5-ITGB1-CAL4A3 complex	0.056	0.359	3	ITGA5	82	
CPLX3	complexin 3	SNARE complex (VAMP2, SNAP25, STX1a, STX3, CPLX1, CPLX3, CPLX4)	0.478	0.515	6	SNAP25	6	Asphyxia neonatorum (P=0.013)
DAG1	dystroglycan 1	UTM-SGCE-DAG1-CAV1-NOS3 complex	0.251	0.604	6	CAV3	33	Muscle tissue disease (P=0.0023)
DLG4	discs large MAGUK scaffold protein 4	DLG4-DLGAP1-SHANK3 complex	0	0.51	6	FYN	18	Schizophrenia 4 (P=0.036)
EPN1	epsin 1	RalBP1-CCNB1-AP2A-NUMB-EPN1 complex	0.098	0.453	5	CCNB1	6	
ESPL1	extra spindle pole bodies like 1, separase	ESPL1-CDC2 complex	-0.372	0.1	3	PTTG1	55	Spindle cell cancer (P=0.03)
F7	coagulation factor VII	Factor-Xa-TFPI-factor-VIIa-tissue factor complex	0	0.479	4	F10	27	Factor VII deficiency (P=0.0014)
GNB1	G protein subunit beta 1	G protein complex (BTK, GNG1, GNG2)	0.195	0.665	28	PTH1R	53	
GP1BA	glycoprotein Ib platelet subunit alpha	ITGA2b-ITGB3-CD9-GP1b-CD47 complex	0.262	0.523	6	VWF	12	Blood platelet disease (P=0.0006)
GTF2F1	general transcription factor IIF subunit 1	RNA polymerase II complex, (CDK8 complex)	0.18	0.644	44	GTF2F2	17	
HSPA1A	heat shock protein family A (Hsp70) member 1A	HSP70-BAG5-PARK2 complex	0.182	0.48	42	HSPA8	96	Ischemia (P=0.017)
HUS1B	HUS1 checkpoint clamp component B	9b-1b-1 complex	0	0.57	2	RAD1	0	Testicular Leydig cell tumor (P=0.032)
IDH3B	isocitrate dehydrogenase (NAD(+)) 3 non-catalytic subunit beta	Isocitrate dehydrogenase [NAD], mitochondrial	0.158	0.431	3	IDH3A	37	
IRAK2	interleukin 1 receptor associated kinase 2	IRAK1-IRAK2 complex	0.227	0.484	2	(Self)	33	Lymphocytic colitis (P=0.018)
ITGA2B	integrin subunit alpha 2b	ITGA2b-ITGB3-CD9-GP1b-CD47 complex	0.204	0.48	12	TGM2	30	Blood platelet disease (P=0.00038)
KCNQ2	potassium voltage-gated channel subfamily Q member 2	KCNQ2-KCNQ3 complex	0.007	0.331	2	KCNQ3	13	Epilepsy (P=0.0012)
LAMA5	laminin subunit alpha 5	ITGA6-ITGB4-LAMA5 complex	0.192	0.466	6	ITGB4	47	Placental infarction (P=0.024)
LRP5	LDL receptor related protein 5	Norrin receptor complex	0.121	0.442	3	FZD4	26	
MPP3	MAGUK p55 scaffold protein 3	CADM1-4.1B-MPP3 complex	-0.069	0.685	14	ITGB2	20	
MRPL53	mitochondrial ribosomal protein L53	39S ribosomal subunit, mitochondrial	0.21	0.654	78	MRPL12	4	
NDUFAF1	NADH:ubiquinone oxidoreductase complex assembly factor 1	Ecsit complex (ECSIT, MT-CO2, GAPDH, TRAF6, NDUFAF1)	0.149	0.443	12	ECSIT	14	Protein deficiency (P=0.0038)
NEDD8	NEDD8 ubiquitin like modifier	Ubiquitin E3 ligase (CDC34, NEDD8, BTRC, CUL1, SKP1A, RBX1)	0.087	0.343	7	SKP1	22	
NGDN	neuroguidin	AATF-NGDN-NOL10 complex	0.249	0.538	3	NOL10	9	
NPHP4	nephrocystin 4	DVL2-INVS-NPHP4-RPGRIP1L complex	0	0.36	8	NPHP1	8	Kidney disease (P=0.0011)
PATJ	PATJ crumbs cell polarity complex component	CRB1-MPP5-INADL complex	0.27	0.562	11	(Self)	26	
PCCB	propionyl-CoA carboxylase subunit beta	Propionyl-CoA carboxylase	-0.008	0.483	2	PCCA	30	Propionic acidemia (P=0.0048)
PHF21B	PHD finger protein 21B	LSD1 complex	0	0.342	13	HMG20A	7	Colon squamous cell carcinoma (P=0.03)
PHKG1	phosphorylase kinase catalytic subunit gamma 1	Phosphorylase kinase complex	-0.629	-0.17	3	(Self)	8	Alzheimer's disease (P=0.041)
PPP3R1	protein phosphatase 3 regulatory subunit B, alpha	Calcineurin-FKBP12 complex	0.263	0.544	3	PPP3CA	21	
PSMB5	proteasome 20S subunit beta 5	26S proteasome	0.022	0.53	37	PSME2	43	Multiple myeloma (P=0.037)
PSMB6	proteasome 20S subunit beta 6	26S proteasome	0	0.53	37	PSME2	31	
PSMC3IP	PSMC3 interacting protein	TBPIP/HOP2-MND1 complex	0.294	0.551	2	MND1	42	Recurrent ovarian germ cell neoplasm (P=0.017)
PSMG2	proteasome assembly chaperone 2	PAC1-PAC2 complex	0.211	0.472	2	PSMG1	13	Herpes simplex (P=0.03)
REV1	REV1 DNA directed polymerase	Rev1-Rev3-Rev7-Polkappa complex	0	0.5	4	REV3L	14	Brucellosis (P=0.0008)
RPA3	replication protein A3	RPA complex	0.211	0.537	20	RPA2	45	Combined thymoma (P=0.048)
RPGRIP1L	RPGRIP1 like	DVL2-INVS-NPHP4-RPGRIP1L complex	-0.051	0.289	4	INVS	6	N syndrome (P=0.0079)
RPS29	ribosomal protein S29	Nop56p-associated pre-rRNA complex	0.06	0.455	120	TUBB1	17	
SERPINA1	serpin family A member 1	SERPINA1-CTSG complex SERPINA1-ELA2 complex	0.201	0.467	3	CTSG	73	Lung disease (P=0.000008)
SERPIND1	serpin family D member 1	MLL-HCF complex	0	0.581	7	MEN1	17	Cervical cancer (P=0.00094)
SMCR8	SMCR8-C9orf72 complex subunit	WDR41-(C9orf72-SMCR8)-(FIP200-ULK1-ATG13-ATG101) complex	0.256	0.571	7	C9orf72	9	Amyotrophic lateral sclerosis (P=0.026)
TEN1	TEN1 subunit of CST complex	CST complex	0	0.383	3	STN1	5	Coats disease (P=0.016)
TNIP2	TNFAIP3 interacting protein 2	TNF-alpha/NF-kappa B signaling complex	0.161	0.629	25	MAP3K8	17	Malignant histiocytic disease (P=0.022)
TTR	transthyretin	TTR-RBP complex	0.084	0.414	2	RBP4	21	Amyloidosis (P=0.0000081)
UBA52	ubiquitin A-52 residue ribosomal protein fusion product 1	60S ribosomal subunit, cytoplasmic Ribosome, cytoplasmic	0	0.537	79	RPS9	20	Iridocyclitis (P=0.026)
YBX3	Y-box binding protein 3	H2AX complex II	0.141	0.424	5	NPM1	55	Lyme disease (P=0.0012)

Table 1: Top 50 proteins whose abundance is under substantial influence from non-cognate transcripts. Columns 1 and 2: gene names. Columns 3 show the representative CORUM complex the gene of interest belongs to. Columns 4 and 5 denote the increase in prediction performance between elastic net single feature (self transcript) vs. CORUM feature sets. Column 6 shows the number of transcripts used to predict the protein level of the gene of interest in the CORUM feature set. Column 7 shows the top trans-locus contributor to the protein level of the gene of interest, ranked by absolute coefficients in the elastic net model. Proteins whose own transcripts are the top predictors are marked with (self). Column 8 denotes the number of MSigDB C2 CGP (chemical and genetic perturbation) gene sets in which the gene appears. Column 9 denotes the top significantly associated Disease Ontology term with the gene of interest in the literature.

Network representation of the interdependencies of protein level regulations

We next examined whether the interdependency of proteins and trans locus transcripts may be used to infer potential novel regulatory drivers. To do so, we constructed graphical models to visualize the connections between each protein and the transcriptome features that contribute to its predicted level, followed by topological analysis to find hub nodes and extract network patterns. To limit scope, we generated graphs to proteins whose prediction improved in the CPTAC_8 data set following the inclusion of the CORUM or STRING features ($\Delta r \geq 0.25$) A directed graph is then generated using a list of edges that connect feature variables (transcripts) to target variables (proteins), with the edge weights calculated using a function of the elastic net coefficients and random forest feature importance of the models for both the CORUM (**Supplementary Table S6**) and the STRING (**Supplementary Table S7**) feature sets. The resulting network is partitioned into connected components of transcript-protein relationships. The subgraph that corresponds to the relationship between PCCA and PCCB as discussed above shows the PCCB node contributing (as transcript) to its own protein level weakly, and the PCCA node contributing strongly to the PCCB node (as protein) (**Figure 6A**). The relationship of CMC1 with other cytochrome c oxidase complex subunits, as described above, is represented in a wheel and spokes pattern where multiple transcripts (orange nodes) contribute either positively or negatively to the CMC1 protein level, with MT-CO1 having the strongest positive contributions (**Figure 6B**). Another member of the MITRAC module COA3 to which CMC1 binds likewise exerts a positive influence on the protein level of CMC1, consistent with the possibility that interdependent protein and transcript relationships may reflect information related to protein complex assembly sequences. Larger subgraphs represent instances where two or more protein target nodes are connected by shared transcript source nodes. Due to the method by which the graphs are generated, protein hubs (with high in-degrees) are more prevalent, but transcripts with high outflow activities are also seen that

connect two or more protein nodes. A subgraph in **Figure 6c** predicts that the transcript levels of SH3KBP1 and CD2AP both contribute to the predicted protein levels of multiple proteins involved in the RICH1/AMOT polarity complex and the PI4K2A-WASH complex, and the protein level of WASHC1 is influenced by genes in the CCC-Wash (WASH1, FAM21C) complex and the PI4K2A-WASH complex (**Figure 6C**).

We highlight an instance where mRNA-protein relationships may be examined to generate new hypotheses. The mitochondrial ribosome is represented in a highly connected subgraph where the abundances of a majority of complex subunit proteins are each contingent upon multiple transcripts (**Figure 6D**). Using topological analysis based on the hierarchical community detection framework (HiDeF) method [33], we showed that this subgraph preserves the expected hierarchical relationship between the 28S small and 39S large subunits in forming the 55S mitochondrial ribosome, suggesting the preferential connections within the proteogenomic graphs broadly recapitulate subcomplex assembly (**Figure 6E**). We then extracted the hub nodes from the subgraph using the betweenness centrality algorithm , which revealed mitochondrial serine beta-lactamase-like protein (LACTB) to be an upstream gene that is highly connected within the subgraph and exerts a regulatory effect on both small and large mitochondrial ribosome subunit components. Feature importance analysis further suggests the *LACTB* transcript to have a negative impact on multiple ribosomal mitochondrial proteins (**Figure 6F**). Of note, *LACTB* was previously identified as a new subunit (MRP-L56) of the mitochondrial ribosomal large subunit isolated from sucrose gradients [58]. Subsequent work however has established the LACTB protein as a filament-forming protein that is localized to the intermembrane space instead away from the mitochondrial ribosomes [59] that possesses in vitro protease activity [60] and that acts as a tumor suppressor by maintaining post-mitotic differentiation states [60]. The physiological function and antiproliferative mechanism of LACTB are incompletely understood, although induced LACTB expression in cancer cells was found to affect mitochondrial phospholipid metabolism [60]. The analysis here therefore raises the possibility that *LACTB* may also affect oxidative phosphorylation in differentiated cells by regulating mitochondrial ribosome biogenesis and protein synthesis, which may be validated experimentally.

A similar hub node analysis from the largest subgraph from the STRING feature set (**Supplemental Table S8**) showed chromogranin A and B (CHGA and CHGB) as inward hubs (i.e., proteins associated with many transcripts) associated with secreted peptides including CST3, SERPINC1, MFGE8, SERPIND1, and others (**Supplementary Figure S9**), which is consistent with the known roles of these acidic glycoproteins as the primary constituents of secretory granules that can regulate their rate of their formation[52]. Taken together,

the topological analysis shown here suggests mRNA-protein networks may be useful for generating new hypothesis on regulatory drivers from large proteogenomics data.

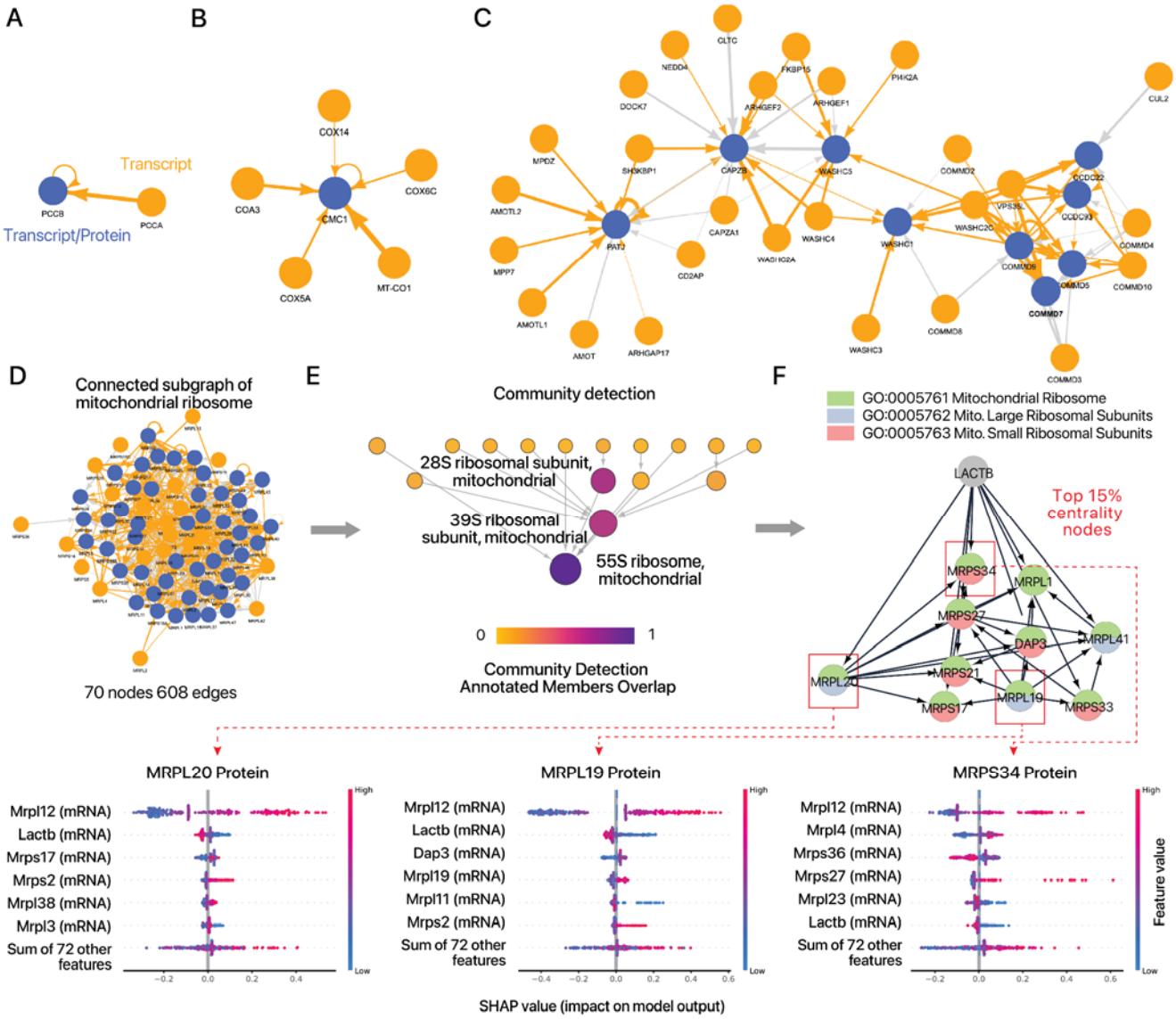


Figure 6: Directed graphs of protein and transcript interrelationships identify candidate regulatory genes. A-C. Examples of directed graphs constructed from genome-wide relationships of transcript-predicted proteins, containing members of **A.** the propionyl-CoA carboxylase complex; **B.** the cytochrome c oxidase, mitochondrial complex; **C.** the PI4K2A-WASH complex, the RICH1/AMOT polarity complex, and others. In each subgraph, orange nodes have outflow

edges only (i.e., they are contributing transcripts in the prediction models). Blue nodes are nodes that are connected to other nodes via at least one inflow edge (i.e., they represent proteins, and optionally also transcripts if they also have outward edges). Orange edges represent positive coefficients of the transcripts to the target proteins in the elastic net models; gray edges represent negative coefficients. All edges are directed from transcript to protein, and the widths of the edges are scaled by the weight. **D.** A highly connected subgraph of mitochondrial ribosome subunits containing 73 nodes and 834 edges. **E.** Persistent community detection and network representation of preferential node connections, showing a hierarchical relationship between the 28S and 39S subcomplex with the assembled 55S mitochondrial ribosome. **F.** Network representation of hub nodes defined as 15% of nodes ranked by betweenness centrality, which predicts a potential role of LACTB as a critical hub that lies upstream of multiple large and small mitochondrial ribosomal protein subunits. Node colors represent the pie chart diagram of the corresponding GO biological process described in the table. SHAP values of three proteins (MRPL20, MRPL19, MRPS34) are highlighted showing top model contributors.

Discussion

Widespread contributions of trans locus transcripts to protein level

Recent work has applied various machine learning models to the computational task of predicting across-sample protein levels using transcriptomics data, but investigation into the biological factors that uncouple transcriptome and proteome data have remained limited, noting only that predictability may differ across broad functional categories, e.g., essential genes or metabolic genes may be less predictable. The results here suggest that the transcript levels of interacting partners had an outsized contribution to the abundance of proteins of interest. The notion that protein-protein interaction can influence protein abundance post-transcriptionally is not new, as it is known that supernumerary subunits of protein complexes can be removed through protein degradation. However, details of genes that show especially poor mRNA and protein abundance correlations and the identities of their protein regulators have remained scarce. The current study provides new evidence from protein prediction models that protein abundance by other transcripts is very common in the proteome, nominating specific interactions involving not only large megadalton sized multi-protein complexes as previously observed, but also smaller stable complexes (e.g., propionyl-CoA carboxylase with two subunits) in the CORUM feature set and potentially more transient interactions documented in the STRING feature set. The common presence of small stable complexes in the proteome greatly expands the repertoire of proteins for which due considerations should be given when directly interpreting transcript level data as representing protein level information. Moreover, it has been suggested that promiscuous protein-protein interactions without established biological function may be a common

occurrence during the co-evolution of functional protein-protein interactions [62]. The occurrence of such stable but non-functional interacting pairs could further increase the scope of trans locus protein regulation which would have implications on the predictability of protein levels from mRNA levels.

The imperfect correlation between protein and mRNA points to orthogonal information that exists between transcript and protein regulations, which underpins the untapped potential for further multi-omics integration to derive new insights [63]. The use of protein correlations to find causal insights for post-transcriptional regulations has previously been explored in other contexts, which looked for anti-correlation between E3 ubiquitin ligases with known proteins or ubiquitination sites of interest, which may control their protein level by virtue of post-translational degradation [18,24]. Extending this intuition towards multi-omics correlation, we used directed graphs generated from the model coefficients and feature importance of trans locus transcripts to represent the interaction patterns between the transcripts and proteins of interest. This in turn nominated a number of hub proteins whose abundance is contributed by multiple transcripts and hub transcripts that regulate multiple proteins, and enabled community detection analysis to find the relationships between biological processes in protein and mRNA correlation networks. As more data continue to become available, we foresee that graphical models will be useful for finding more trans locus regulations of protein levels, such as those that represent known assembly sequential steps or post-transcriptional regulators. These graphical models may be used to generate testable hypotheses or find utility in predicting experimental outcome, barring confounders or reverse causality effects. For instance, in the case of the actin related protein 2/3 complex, one might predict that within a certain concentration range an overexpression of *ACTR3* and *ARPC4* will be more effective in modulating ARPC3 protein levels than augmenting the expression of *ARPC3* itself, which is readily testable by experimentation.

Limitations of the study

Significant differences in protein regulations likely occur in different cell and tissue types. The predictive models here are trained using publicly available CPTAC data from 8 cancer types, which contain transcriptome and proteome data from both tumor and normal adjacent tissue samples. In prior work, we found general concordance between protein and mRNA correlation in CPTAC samples vs. GTEx tissue proteomics data [15]. However, other studies that compared protein and mRNA correlation in tumors and normal adjacent tissues have found higher inter-sample correlation in tumor samples [21,22], which may be attributable to the increased translation rates in cancer. Hence, additional discordant cases between proteins and mRNA likely remain to be discovered that are omitted here. Sample difference may also explain the

observation from GTEx proteomics that secreted proteins are associated with protein and mRNA discordance [7] as one might intuitively assume but which is not apparent in the analysis here. The cases used for training the model are not labeled by their cancer type or their tumor vs. normal adjacent tissue designation, which would likely have availed overall predictive performance.

We performed the analysis using expression data from TCGA/CPTAC RNA sequencing experiments through the cptac Python API, which retrieves the final data tables from the flagship CPTAC papers of each individual cancer type [25]. Although each TCGA/CPTAC cancer subtype project follows an overall consistent experimental design and data acquisition strategy, minute differences exist in the processing pipelines used to analyze the RNA sequencing data (e.g., STAR vs. Bowtie2) and gene expression measure (e.g., RPKM vs. FPKM) which could bias gene expression values across cancer types. Likewise, there are subtle differences in the protein expression data in each cancer type (e.g., search engine, isotope tags). As our goal here is not to compare trends across cancer types, we have taken inspiration from the CPTAC Dream Challenge submissions that improved overall predictions by borrowing information across cancer types. However, it is also known that different data sets may present different variability. Future work may employ uniformed processed data to improve performance and reduce bias, such as data from the TCGA PanCanAtlas which processed all TCGA cancer RNA sequencing data uniformly [64], or compare batch correction strategies. To our knowledge, uniformly processed protein data are not yet available at the time of writing. Other machine learning and deep learning algorithms can likely further boost protein level predictions. We have limited our scope here to comparisons of different feature sets and previously employed algorithms, and to result interpretation.

Finally, pitfalls must be considered when attempting to interpret predictive models in search of mechanistic insights. Feature importance in the predictive models represent correlation rather than causality, and hence interpretations can be confounded by independent confounders, multicollinearity, and reverse causality. In simple cases such as PCCB–PCCA, given the emphasis on prior feature selections that prioritize known protein-protein interaction partners, the logical interpretation would be to assume the hierarchical nature of gene regulations where transcript levels are more likely to affect protein abundance than the opposite, but this assumption becomes more tenuous as feature sets expand and the number of associations increase, and in cases such as transcription factor proteins whose abundance can affect transcript levels across samples. Future work in this area may employ more sophisticated causal inference methods to identify regulatory modalities.

Conclusion

In summary, this study compared predictive models of protein levels using different transcript feature sets, and provided biological interpretations of the results by highlighting trans locus transcripts with substantial contributions to protein levels. Although the concept of proteins being modulated by their interacting partners is not new, to our knowledge the prevalence of these regulations have not been examined at a proteome wide level and using recently available proteogenomics data sets. The analysis here therefore reveals new details into the gene identity and modality of trans regulation of protein levels, and gives support to further development of prior transcript feature selection strategies to optimize protein prediction tasks. The results show that the transcript levels of protein-protein interaction partners can broadly influence protein abundance in a tissue, which has implications on the interpretations of transcriptomics data and on understanding the architecture of proteome composition regulations. With further refinement of feature selection and feature engineering methods and the availability of large data sets, we foresee that similar approaches to those shown here will provide valuable new insights into post-transcriptional mechanisms of protein regulations.

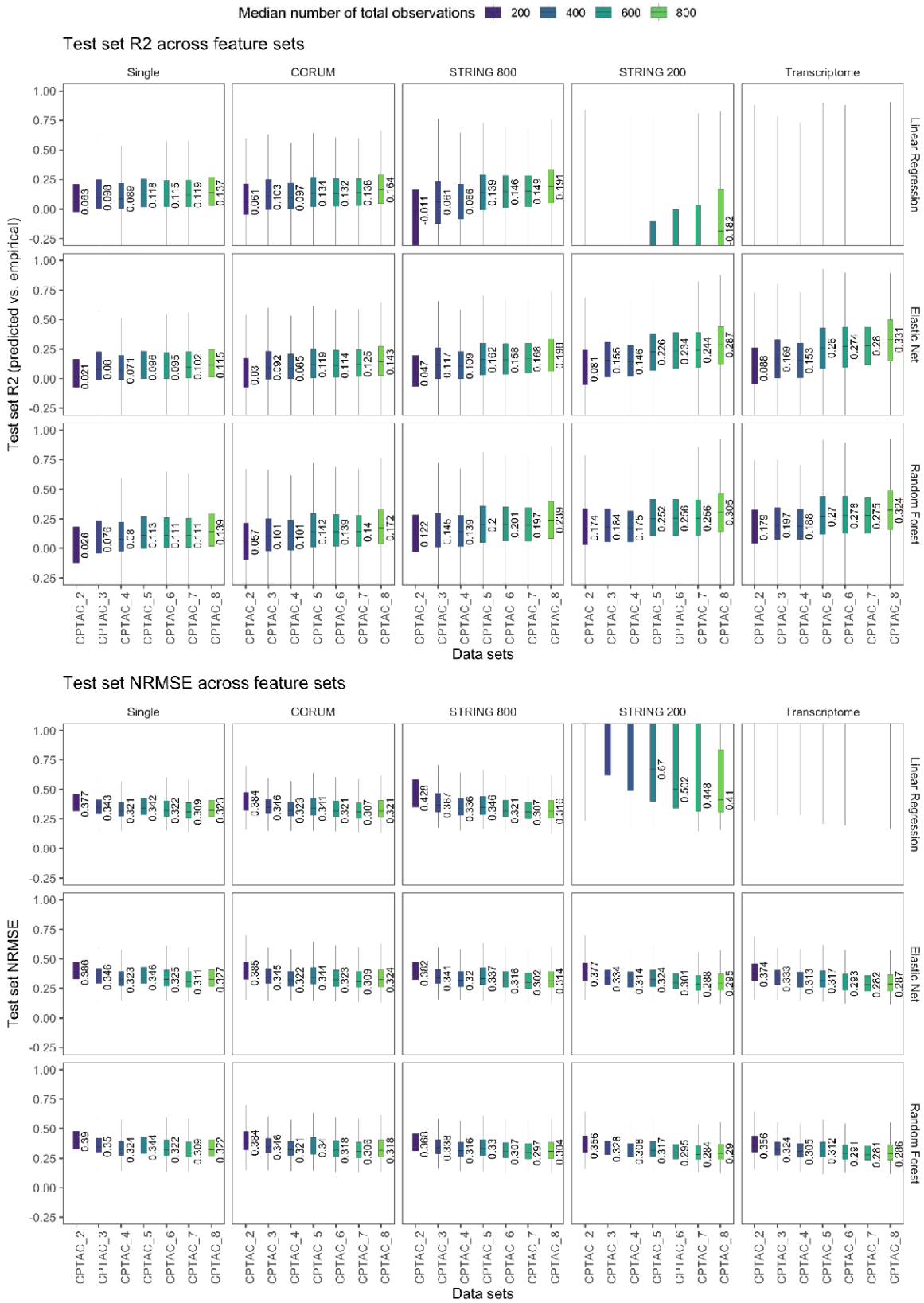
Author Contributions

EL conceptualized the project. HS and EL wrote software code. HS, MJL, and EL performed data analysis. HS, MJL, JC, RC, MPL, and EL drafted the manuscript. MPL and EL revised the manuscript.

Acknowledgments

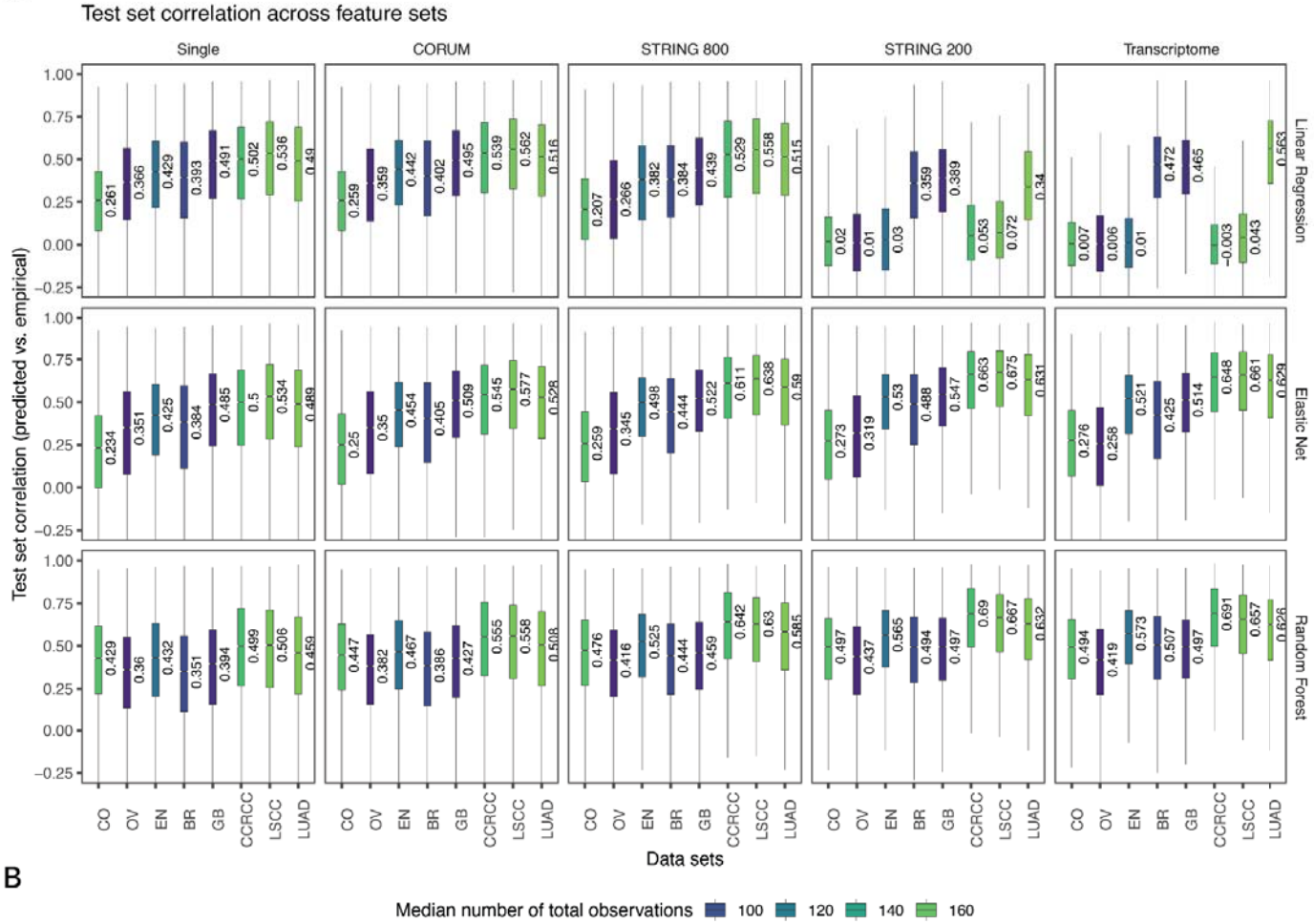
This work was supported in part by NIH Office of the Director award R03-OD032666; NIH/NHLBI award R00-HL144829 to EL; and NIH/NHLBI awards R00-HL127302, R01-HL141278, and the Consortium for Fibrosis Research & Translation funds at the University of Colorado to ML.

Supplementary Figures

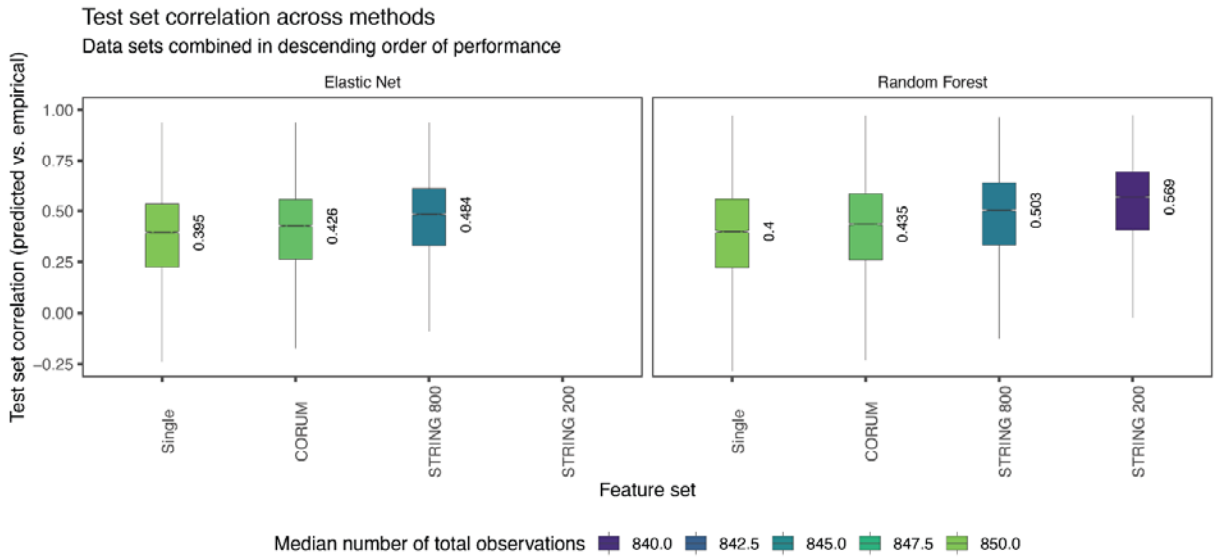

Median number of total observations 200 400 600 800
A
Test set R2 across feature sets
Single
CORUM
STRING 800
STRING 200
Transcriptome
Linear Regression
Elastic Net
Random Forest
Test set R2 (predicted vs. empirical)
Data sets
B
Test set NRMSE across feature sets
Test set NRMSE
Data sets

Supplementary Figure S1: Model performance measured by additional metrics. **A.** Box plots of test set normalized root mean square error (NRMSE) between the transcript-predicted and actual protein level for each protein are shown across five feature sets (column: single/self transcript, CORUM interactors, STRING high-confidence associated proteins; STRING low-confidence associated proteins, and all transcripts) and three algorithms (multiple linear regression, elastic net, and random forest). In each plot, x axis denotes the number of CPTAC data set used to train the models box: interquartile range; whiskers: +/– 1.5 IQR. **B.** As above, but for test set goodness-of-fit (R^2).

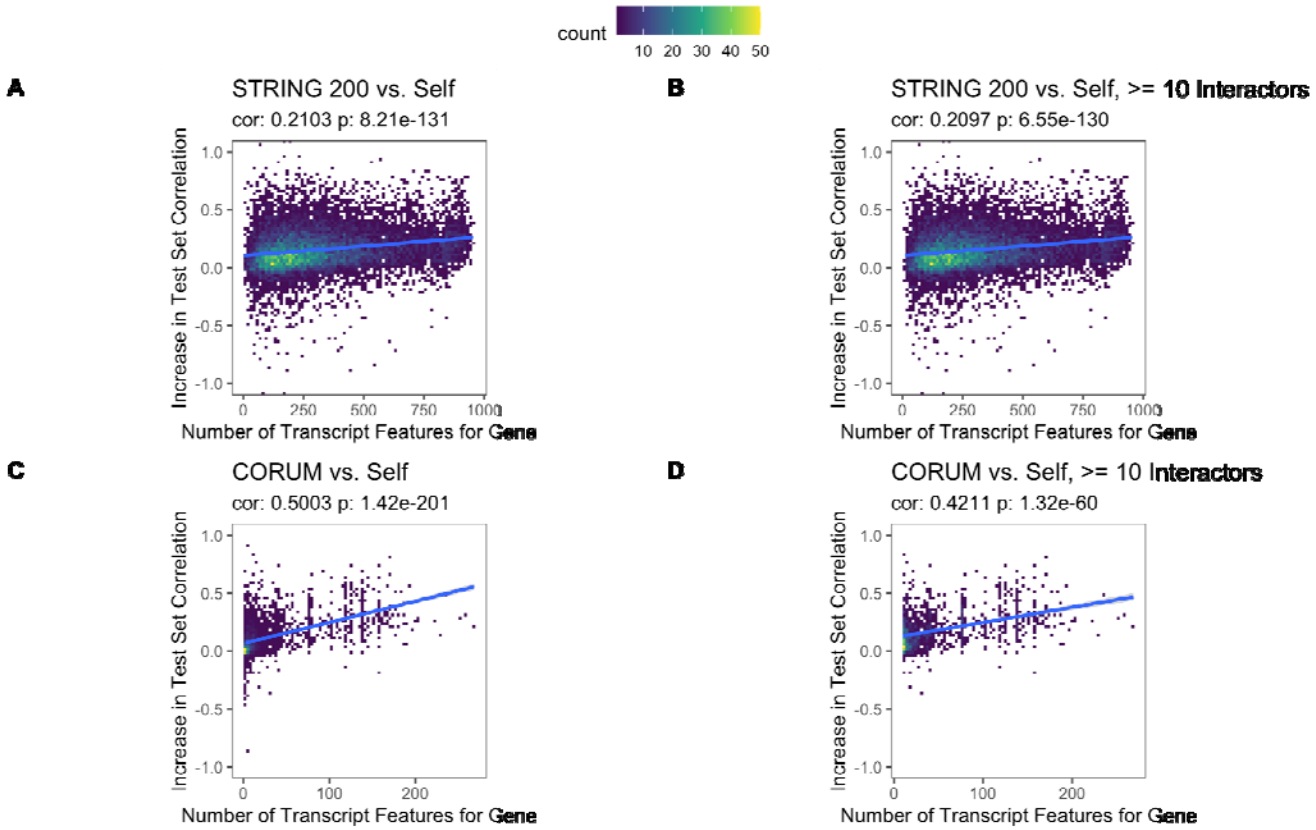
A


Test set correlation across feature sets
Single
CORUM
STRING 800
STRING 200
Transcriptome
Linear Regression
Elastic Net
Random Forest
Test set correlation (predicted vs. empirical)
CO
OV
EN
BR
GB
CCRCC
LSCC
LUAD
Data sets
Median number of total observations
100
120
140
160

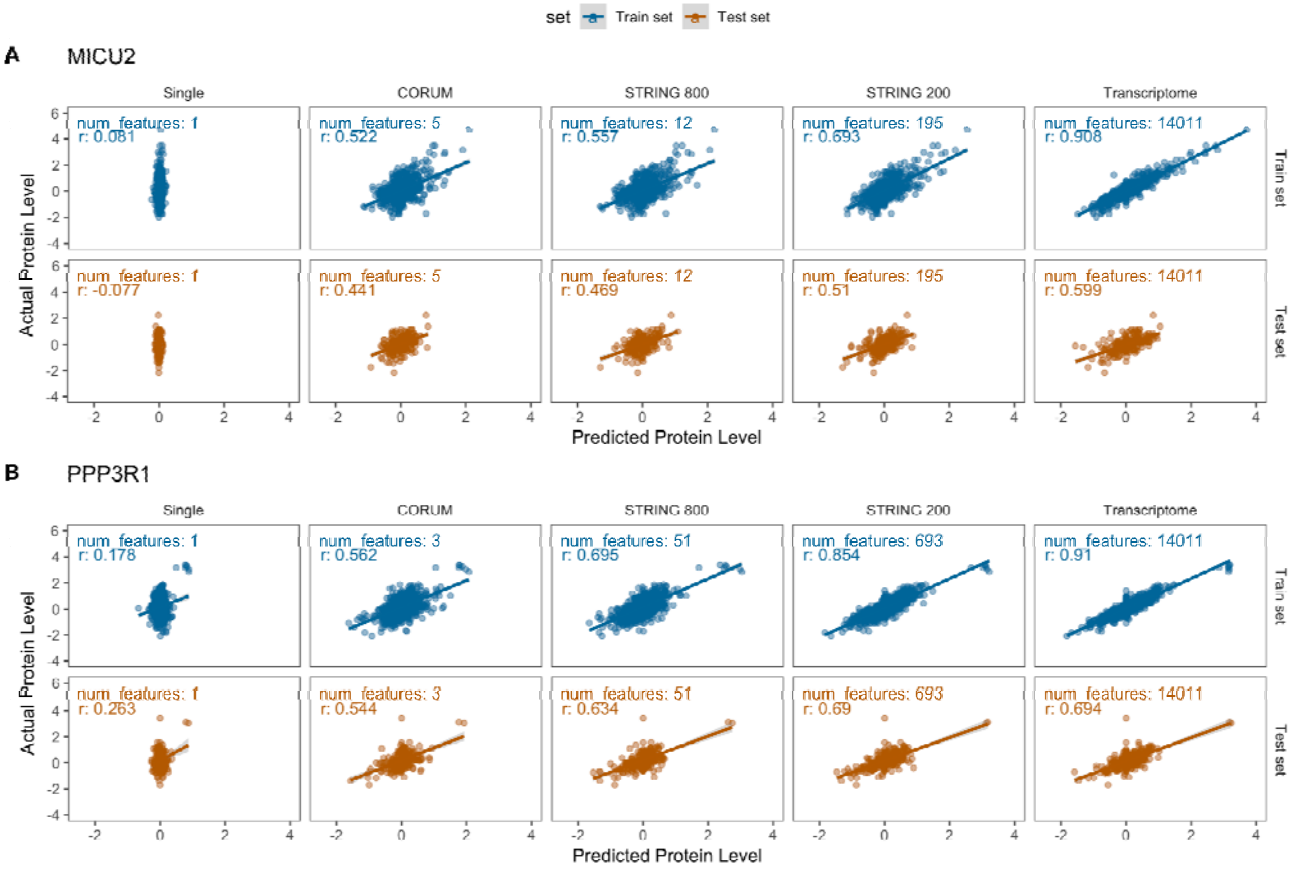
B


Test set correlation across methods
Data sets combined in descending order of performance
Elastic Net
Random Forest
Test set correlation (predicted vs. empirical)
Single
CORUM
STRING 800
STRING 200
Feature set
Median number of total observations
840.0
842.5
845.0
847.5
850.0

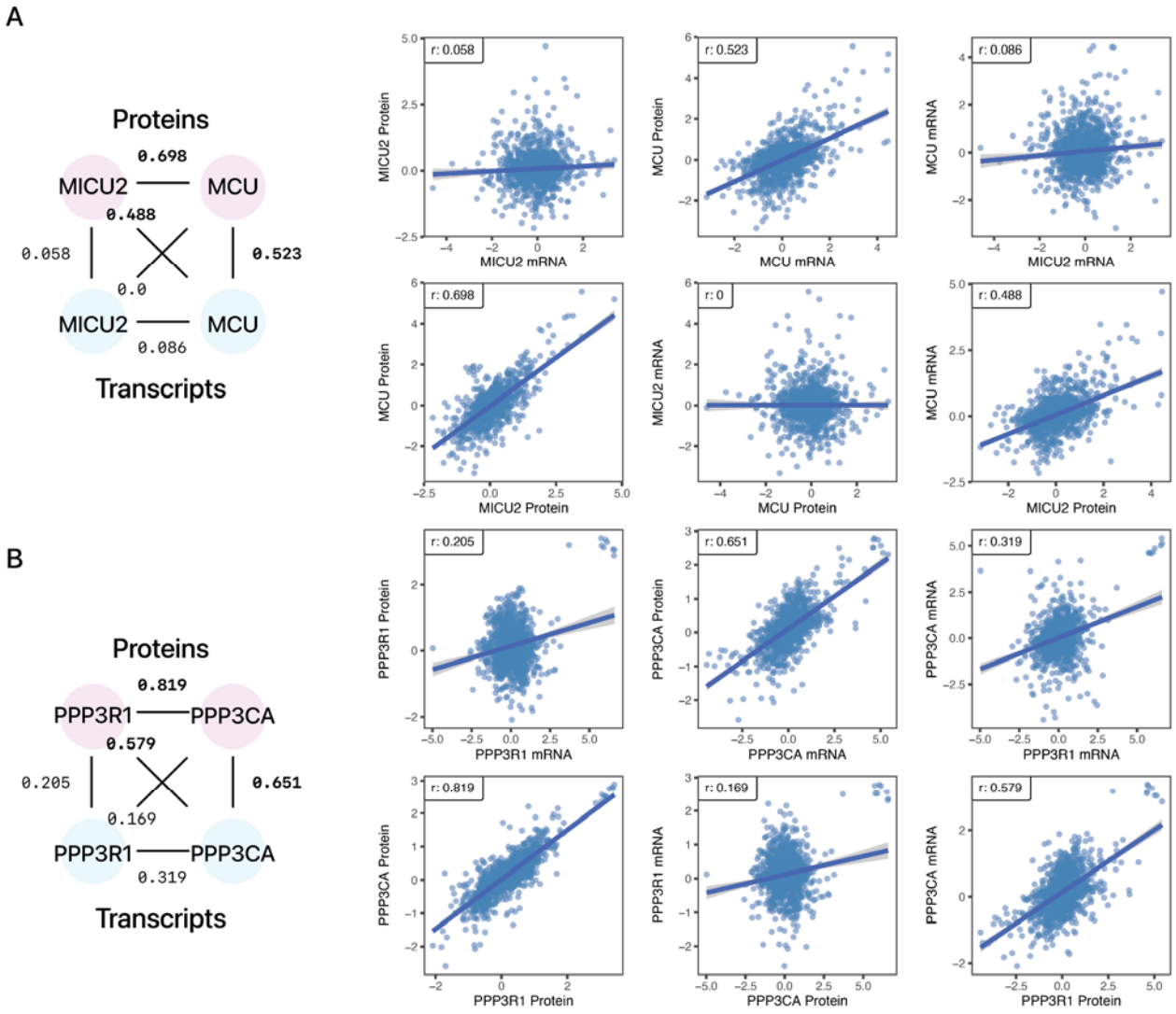
Supplementary Figure S2: A. Model performance in single cancer data sets. Box plots of test set correlation coefficients (r) between the transcript-predicted and actual protein level for each protein are shown across five feature sets (column: single/self transcript, CORUM interactors, STRING high-confidence associated proteins; STRING low-confidence associated proteins, and all transcripts) and three algorithms (multiple linear regression, elastic net, and random forest). In each plot, x axis denotes the CPTAC cancer type study used to train the models; box: interquartile range; whiskers: +/– 1.5 IQR. **B.** Model performance when the 8 data sets were combined in the order of decreasing single data set performance.



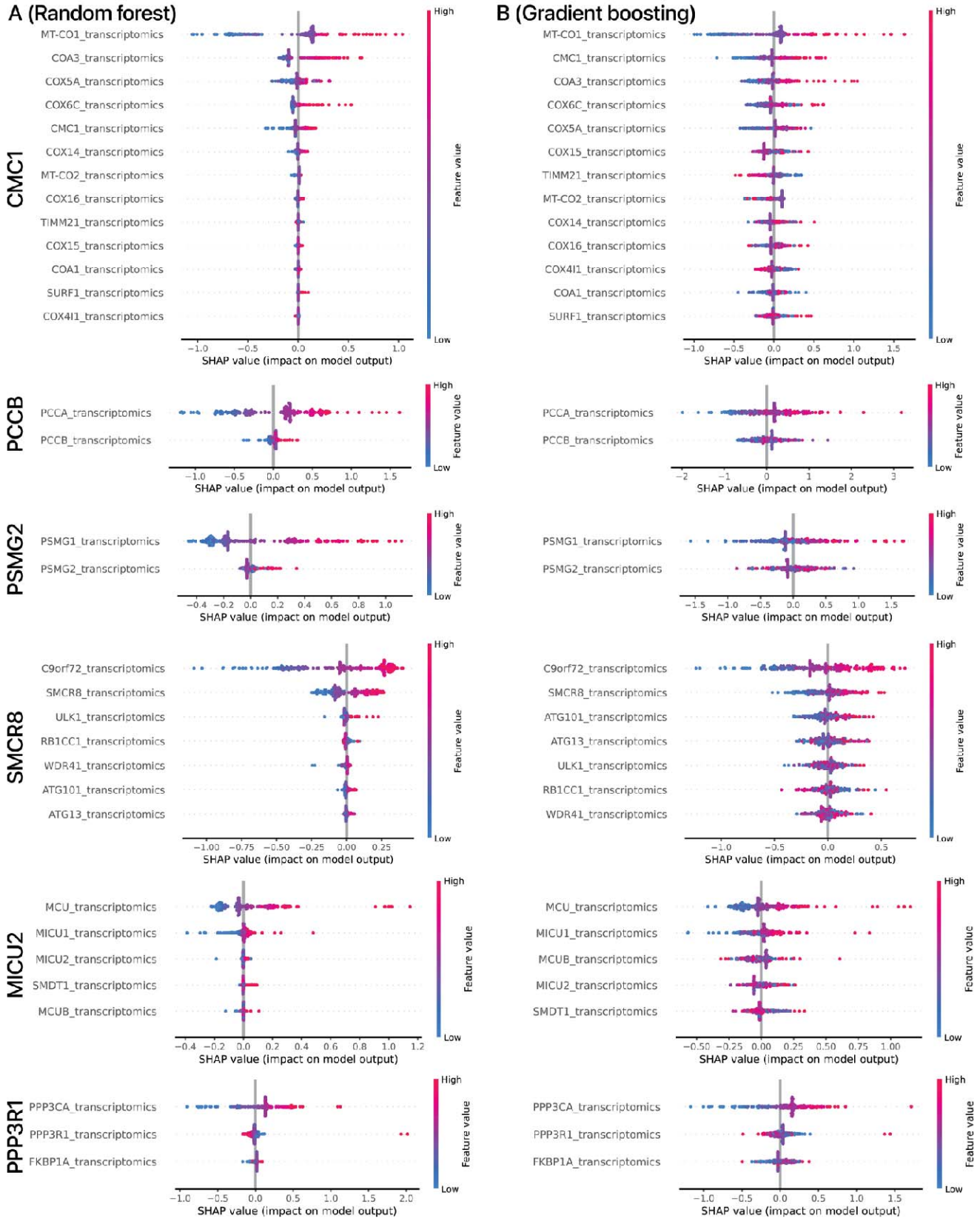
Supplementary Figure S3. Improvements to protein prediction from the incorporation of additional transcript features. A–B. Scatterplot showing a significant linear relationship between the number of low-confidence interactors a protein has as annotated in STRING vs. the increase in test set correlation between predicted vs. actual protein levels in the Elastic Net models over the self-transcript feature set (correlation test p: 8.2e–131 for all proteins; 6.5e–130 for proteins with 10 or more interactors). **C–D.** As above, but for the CORUM feature set.



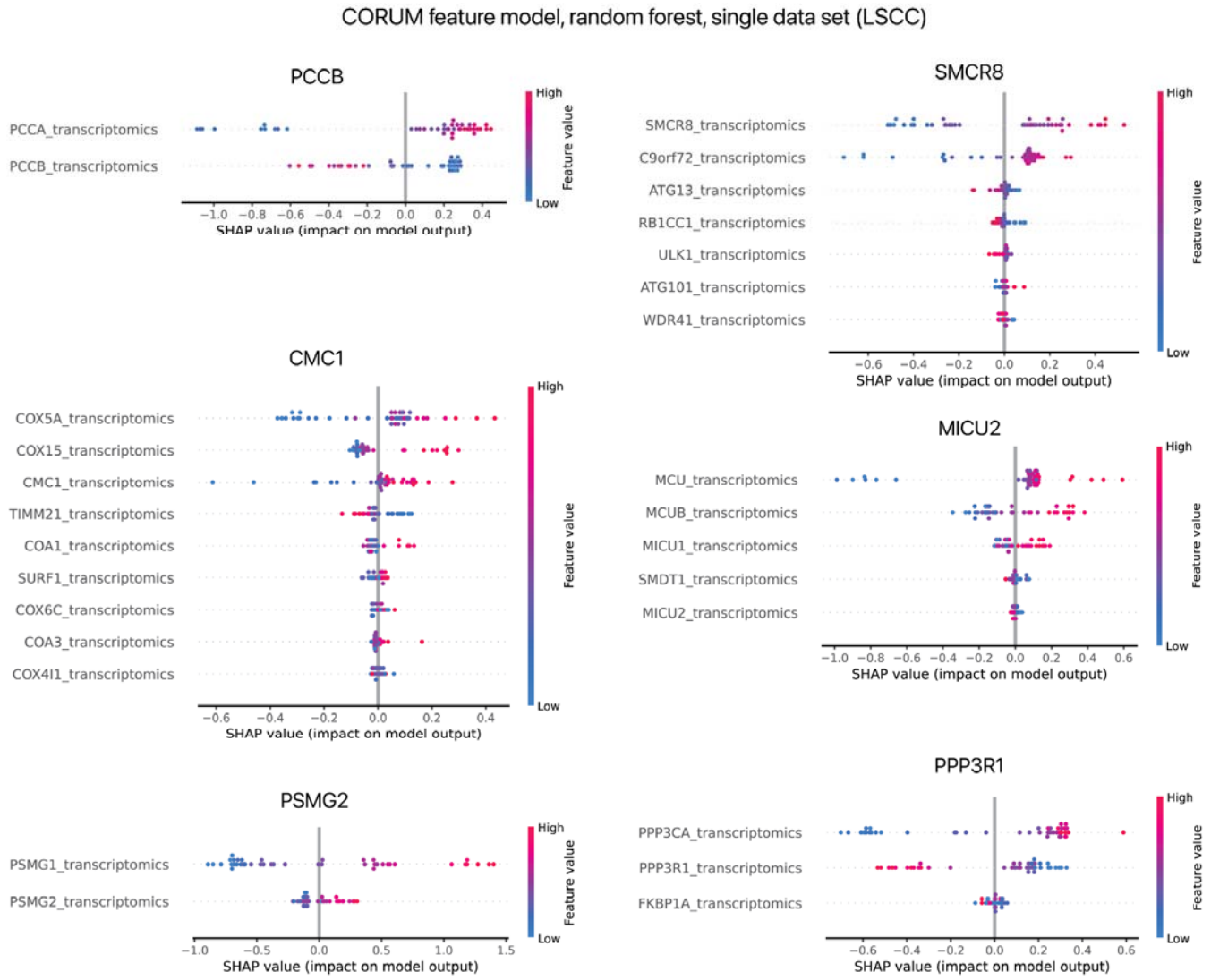
Supplementary Figure S4: Additional proteins with improved predicted levels after inclusion of additional transcript features. Two additional highlighted proteins with substantial predictability from transcriptome data upon the inclusion of additional features are shown: **A.** MICU2, **B.** PPP3R1. For each protein, the transcript-trained prediction of protein level is plotted on the x axis and the actual protein level is plotted on the y axis. Blue: train set, brown: test set. Columns denote the transcript feature set used to train the model. The number of features used to train the model in each feature set is shown inside each plot. r: Correlation coefficient.



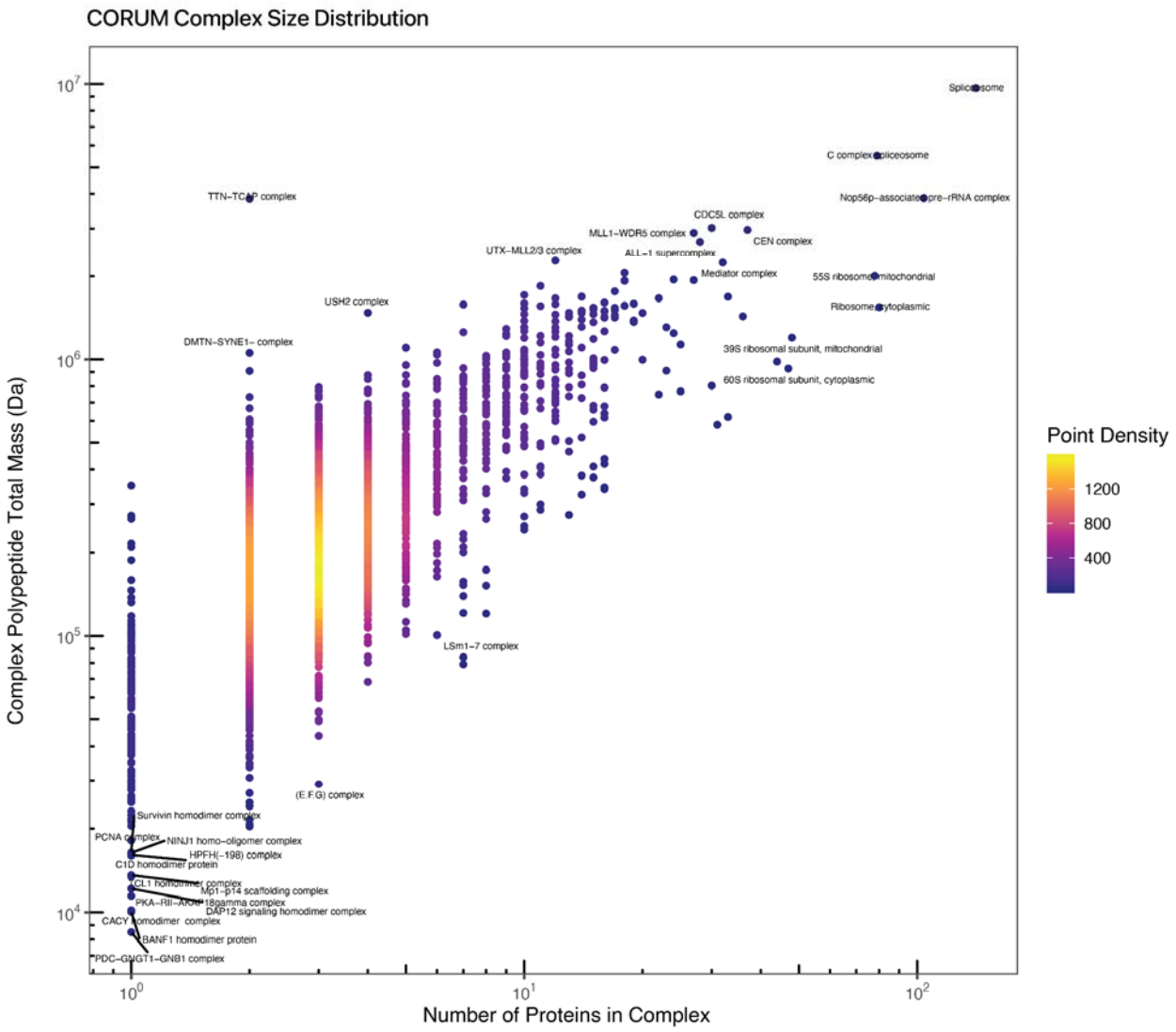
Supplementary Figure S5: Cross-omics co-expression of MICU2 and PPP3R1 with functionally associated proteins. Two examples of proteins whose abundance is better explained by another transcript are shown. **A.** MICU2 protein level is predicted by MCU transcript but not its own transcript. **B.** PPP3R1 protein level is explained by PPP3CA transcript level but not its own transcript. Substantial correlations across transcripts and proteins (≥ 0.4) are bolded.


A (Random forest)
B (Gradient boosting)
CMC1
MT-CO1_transcriptomics
COA3_transcriptomics
COX5A_transcriptomics
COX6C_transcriptomics
CMC1_transcriptomics
COX14_transcriptomics
MT-CO2_transcriptomics
COX16_transcriptomics
TIMM21_transcriptomics
COX15_transcriptomics
COA1_transcriptomics
SURF1_transcriptomics
COX4I1_transcriptomics
CMC1_transcriptomics
COX15_transcriptomics
PCCB
PCCA_transcriptomics
PCCB_transcriptomics
PSMG2
PSMG1_transcriptomics
PSMG2_transcriptomics
SMCR8
C9orf72_transcriptomics
SMCR8_transcriptomics
ULK1_transcriptomics
RB1CC1_transcriptomics
WDR41_transcriptomics
ATG101_transcriptomics
ATG13_transcriptomics
MICU2
MCU_transcriptomics
MICU1_transcriptomics
MICU2_transcriptomics
SMDT1_transcriptomics
MCUB_transcriptomics
PPP3R1
PPP3CA_transcriptomics
PPP3R1_transcriptomics
FKBP1A_transcriptomics
SHAP value (impact on model output)
Feature value
High
Low

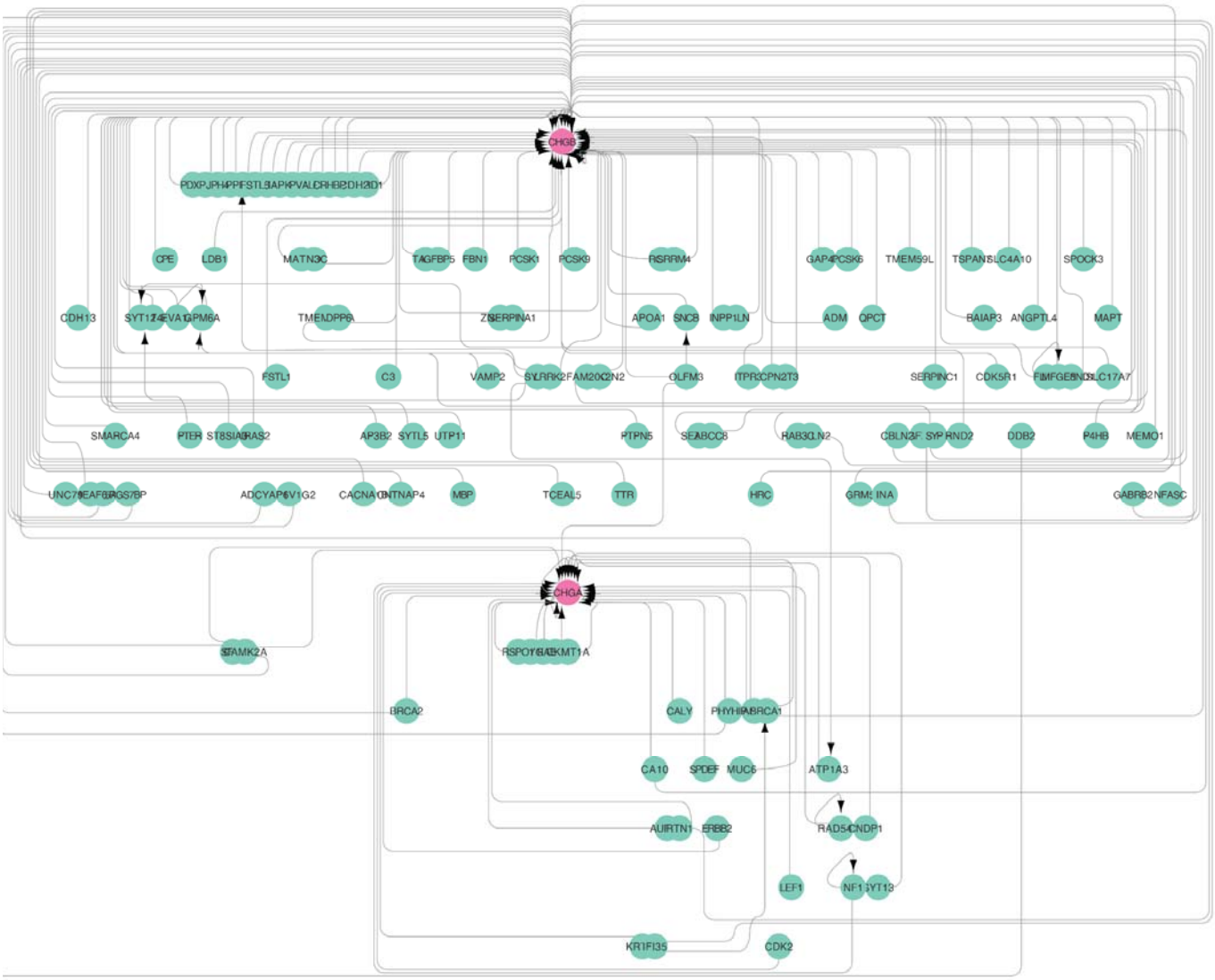
Supplementary Figure S6: SHAP interpretation of feature importance in **A.** random forest and **B.** gradient boosting model output from the CPTAC_8 CORUM feature set. The SHAP values of top transcript features and their impact on model output are shown for each of six proteins highlighted in the text (from top to bottom): PCCB, CMC1, PSMG2, SMCR8, MICU2, PPP3R1.



Supplementary Figure S7: SHAP interpretation of feature importance in the CORUM feature set, random forest model, of a single data set (LSCC), showing largely conserved observations among the highlighted proteins PCCB, CMC1, PSMG2, SMCR8, MICU2, PPP3R1. Consistent with the analysis of the combined CPTAC_8 data set, five of the six proteins are not best predicted by their cognate transcripts and all but one (MT-CO1) of the top trans locus transcripts are preserved, which was not among the examined candidate features. Only test set data points are shown.



Supplementary Figure S8: Density plot showing the distribution of complex sizes (x:-axis: number of member sin complex; y-axis: log10 total polypeptide molecular weight in complex) in the annotated feature set derived from CORUM v.3.0. Names of select complexes are labeled. The majority of complexes are small with a median of 3 proteins per complex.



Supplementary Figure S9: The most connected subgraph in the STRING feature set contains 6,319 nodes and 11,161 edges, containing proteins belonging to multiple distinct cellular compartments and multi-protein complexes. The hierarchical network diagram shows Chromogranin A and B (CHGA/CHGB) (magenta) and their first-degree inward flow neighbors (green) in the largest subgraph constructed from the STRING data set. The protein level of CHGA and CHGB is associated with the transcript levels of multiple secreted peptide coding genes, consistent with these proteins serve to regulate gene expression in secretion pathways.

Supplementary Tables

Supplementary tables can be accessed on figshare at: https://doi.org/10.6084/m9.figshare.19330541

Supplementary Table S1: Table listing performance metrics (test set correlation coefficients, R^2, NRMSE) of the tested models, feature sets, and data sets.

Supplementary Table S2: Table listing Gene Ontology cellular compartment term enrichment results for proteins that are well predicted by their transcript levels.

Supplementary Table S3: Table listing Gene Ontology cellular compartment term enrichment for proteins that are poorly predicted by their transcript levels.

Supplementary Table S4: Table listing individual genewise model performance metrics (test set correlation coefficients, R^2, NRMSE) in the CPTAC_8 data set using elastic net models on different feature sets.

Supplementary Table S5: Table listing Gene Ontology cellular compartment term enrichment for proteins with substantial improvements in prediction performance upon incorporating STRING features in elastic net models.

Supplementary Table S6: Network edge list for proteins with improved predictions in the CORUM feature set along with the coefficients and feature importances of contributing transcripts.

Supplementary Table S7: Network edge list for proteins with improved predictions in the STRING feature set along with the coefficients and feature importances of contributing transcripts.

upplementary Table S8: Hub genes based on centrality scores in the STRING feature set network.

References

(1) Gygi, S. P.; Rochon, Y.; Franza, B. R.; Aebersold, R. Correlation between Protein and MRNA Abundance in Yeast. *Mol. Cell. Biol.* **1999**, *19* (3), 1720–1730.

https://doi.org/10.1128/MCB.19.3.1720.

(2) Liu, Y.; Beyer, A.; Aebersold, R. On the Dependency of Cellular Protein Levels on MRNA Abundance. *Cell* **2016**, *165* (3), 535–550. https://doi.org/10.1016/j.cell.2016.03.014.

(3) Vogel, C.; Marcotte, E. M. Insights into the Regulation of Protein Abundance from Proteomic and Transcriptomic Analyses. *Nat. Rev. Genet.* **2012**, *13* (4), 227–232. https://doi.org/10.1038/nrg3185.

(4) Franks, A.; Airoldi, E.; Slavov, N. Post-Transcriptional Regulation across Human Tissues. *PLoS Comput. Biol.* **2017**, *13* (5), e1005535. https://doi.org/10.1371/journal.pcbi.1005535.

(5) Upadhya, S. R.; Ryan, C. J. *Experimental Reproducibility Limits the Correlation between MRNA and Protein Abundances in Tumour Proteomic Profiles*; preprint; Systems Biology, 2021. https://doi.org/10.1101/2021.09.22.461108.

(6) An, H.; Harper, J. W. Ribosome Abundance Control Via the Ubiquitin–Proteasome System and Autophagy. *J. Mol. Biol.* **2020**, *432* (1), 170–184. https://doi.org/10.1016/j.jmb.2019.06.001.

(7) Jiang, L.; Wang, M.; Lin, S.; Jian, R.; Li, X.; Chan, J.; Dong, G.; Fang, H.; Robinson, A. E.; GTEx Consortium; Snyder, M. P. A Quantitative Proteome Map of the Human Body. *Cell* **2020**, *183* (1), 269-283.e19. https://doi.org/10.1016/j.cell.2020.08.036.

(8) Gonçalves, E.; Fragoulis, A.; Garcia-Alonso, L.; Cramer, T.; Saez-Rodriguez, J.; Beltrao, P. Widespread Post-Transcriptional Attenuation of Genomic Copy-Number Variation in Cancer. *Cell Syst.* **2017**, *5* (4), 386-398.e4. https://doi.org/10.1016/j.cels.2017.08.013.

(9) Eraslan, B.; Wang, D.; Gusic, M.; Prokisch, H.; Hallström, B. M.; Uhlén, M.; Asplund, A.; Pontén, F.; Wieland, T.; Hopf, T.; Hahne, H.; Kuster, B.; Gagneur, J. Quantification and Discovery of Sequence Determinants of Protein-per-MRNA Amount in 29 Human Tissues. *Mol. Syst. Biol.* **2019**, *15* (2), e8513. https://doi.org/10.15252/msb.20188513.

(10) Wang, D.; Eraslan, B.; Wieland, T.; Hallström, B.; Hopf, T.; Zolg, D. P.; Zecha, J.; Asplund, A.; Li, L.; Meng, C.; Frejno, M.; Schmidt, T.; Schnatbaum, K.; Wilhelm, M.; Ponten, F.; Uhlen, M.; Gagneur, J.; Hahne, H.; Kuster, B. A Deep Proteome and Transcriptome Abundance Atlas of 29 Healthy Human Tissues. *Mol. Syst. Biol.* **2019**, *15* (2). https://doi.org/10.15252/msb.20188503.

(11) Waldbauer, J. R.; Rodrigue, S.; Coleman, M. L.; Chisholm, S. W. Transcriptome and Proteome Dynamics of a Light-Dark Synchronized Bacterial Cell Cycle. *PloS One* **2012**, *7* (8), e43432. https://doi.org/10.1371/journal.pone.0043432.

(12) Yang, M.; Petralia, F.; Li, Z.; Li, H.; Ma, W.; Song, X.; Kim, S.; Lee, H.; Yu, H.; Lee, B.; Bae, S.; Heo, E.; Kaczmarczyk, J.; Stępniak, P.; Warchoł, M.; Yu, T.; Calinawan, A. P.; Boutros, P. C.; Payne, S. H.; Reva, B.; NCI-CPTAC-DREAM Consortium; Boja, E.; Rodriguez, H.; Stolovitzky, G.; Guan, Y.; Kang, J.; Wang, P.; Fenyö, D.; Saez-Rodriguez, J. Community Assessment of the Predictability of Cancer Protein and Phosphoprotein Levels from Genomics and Transcriptomics. *Cell Syst.* **2020**, *11* (2), 186-195.e9. https://doi.org/10.1016/j.cels.2020.06.013.

(13) Eicher, T.; Patt, A.; Kautto, E.; Machiraju, R.; Mathé, E.; Zhang, Y. Challenges in Proteogenomics: A Comparison of Analysis Methods with the Case Study of the DREAM Proteogenomics Sub-Challenge. *BMC Bioinformatics* **2019**, *20* (Suppl 24), 669. https://doi.org/10.1186/s12859-019-3253-z.

(14) Li, H.; Siddiqui, O.; Zhang, H.; Guan, Y. Joint Learning Improves Protein Abundance Prediction in Cancers. *BMC Biol.* **2019**, *17* (1), 107. https://doi.org/10.1186/s12915-019-0730-9.

(15) Han, Y.; Li, L. Z.; Kastury, N. L.; Thomas, C. T.; Lam, M. P. Y.; Lau, E. Transcriptome Features of Striated Muscle Aging and Predictability of Protein Level Changes. *Mol. Omics* **2021**. https://doi.org/10.1039/d1mo00178g.

(16) Xu, W.; He, H.; Guo, Z.; Li, W. Evaluation of Machine Learning Models on Protein Level Inference from Prioritized RNA Features. *Brief. Bioinform.* **2022**, bbac091. https://doi.org/10.1093/bib/bbac091.

(17) Zhang, H.; Liu, T.; Zhang, Z.; Payne, S. H.; Zhang, B.; McDermott, J. E.; Zhou, J.-Y.; Petyuk, V. A.; Chen, L.; Ray, D.; Sun, S.; Yang, F.; Chen, L.; Wang, J.; Shah, P.; Cha, S. W.; Aiyetan, P.; Woo, S.; Tian, Y.; Gritsenko, M. A.; Clauss, T. R.; Choi, C.; Monroe, M. E.; Thomas, S.; Nie, S.; Wu, C.; Moore, R. J.; Yu, K.-H.; Tabb, D. L.; Fenyö, D.; Bafna, V.; Wang, Y.; Rodriguez, H.; Boja, E. S.; Hiltke, T.; Rivers, R. C.; Sokoll, L.; Zhu, H.; Shih, I.-M.; Cope, L.; Pandey, A.; Zhang, B.; Snyder, M. P.; Levine, D. A.; Smith, R. D.; Chan, D. W.; Rodland, K. D.; CPTAC Investigators. Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell* **2016**, *166* (3), 755–765. https://doi.org/10.1016/j.cell.2016.05.069.

(18) Mertins, P.; Mani, D. R.; Ruggles, K. V.; Gillette, M. A.; Clauser, K. R.; Wang, P.; Wang, X.; Qiao, J. W.; Cao, S.; Petralia, F.; Kawaler, E.; Mundt, F.; Krug, K.; Tu, Z.; Lei, J. T.; Gatza, M. L.; Wilkerson, M.; Perou, C. M.; Yellapantula, V.; Huang, K.; Lin, C.; McLellan, M. D.; Yan, P.; Davies, S. R.; Townsend, R. R.; Skates, S. J.; Wang, J.; Zhang, B.; Kinsinger, C. R.; Mesri, M.; Rodriguez, H.; Ding, L.; Paulovich, A. G.; Fenyö, D.; Ellis, M. J.; Carr, S. A.; NCI CPTAC. Proteogenomics Connects Somatic Mutations to Signalling in Breast Cancer. *Nature* **2016**, *534* (7605), 55–62. https://doi.org/10.1038/nature18003.

(19) Dou, Y.; Kawaler, E. A.; Cui Zhou, D.; Gritsenko, M. A.; Huang, C.; Blumenberg, L.; Karpova, A.; Petyuk, V. A.; Savage, S. R.; Satpathy, S.; Liu, W.; Wu, Y.; Tsai, C.-F.; Wen, B.; Li, Z.; Cao, S.; Moon, J.; Shi, Z.; Cornwell, M.; Wyczalkowski, M. A.; Chu, R. K.; Vasaikar, S.; Zhou, H.; Gao, Q.; Moore, R. J.; Li, K.; Sethuraman, S.; Monroe, M. E.; Zhao, R.; Heiman, D.; Krug, K.; Clauser, K.; Kothadia, R.; Maruvka, Y.; Pico, A. R.; Oliphant, A. E.; Hoskins, E. L.; Pugh, S. L.; Beecroft, S. J. I.; Adams, D. W.; Jarman, J. C.; Kong, A.; Chang, H.-Y.; Reva, B.; Liao, Y.; Rykunov, D.; Colaprico, A.; Chen, X. S.; Czekański, A.; Jędryka, M.; Matkowski, R.; Wiznerowicz, M.; Hiltke, T.; Boja, E.; Kinsinger, C. R.; Mesri, M.; Robles, A. I.; Rodriguez, H.; Mutch, D.; Fuh, K.; Ellis, M. J.; DeLair, D.; Thiagarajan, M.; Mani, D. R.; Getz, G.; Noble, M.; Nesvizhskii, A. I.; Wang, P.; Anderson, M. L.; Levine, D. A.; Smith, R. D.; Payne, S. H.; Ruggles, K. V.; Rodland, K. D.; Ding, L.; Zhang, B.; Liu, T.; Fenyö, D.; Agarwal, A.; Anurag, M.; Avtonomov, D.; Birger, C.; Birrer, M. J.; Boca, S. M.; Bocik, W. E.; Borate, U.; Borucki, M.; Burke, M. C.; Cai, S.; Calinawan, A.; Carr, S. A.; Carter, S.; Castro, P.; Cerda, S.; Chaikin, M.; Chan, D. W.; Chan, D.; Charamut, A.; Chen, F.; Chen, J.; Chen, L.; Chen, L. S.; Chesla, D.; Chheda, M. G.; Chinnaiyan, A. M.; Chowdhury, S.; Cieslik, M. P.; Clark, D. J.; Cottingham, S.; Culpepper, H.; Day, J.; De Young, S.; Demir, E.; Dhanasekaran, S. M.; Dhir, R.; Domagalski, M. J.; Dottino, P.; Druker, B.; Duffy, E.; Dyer, M.; Edwards, N. J.; Edwards, R.; Elburn, K.; Field, J. B.; Francis, A.; Gabriel, S.; Geffen, Y.; Geiszler, D.; Gillette, M. A.; Godwin, A. K.; Grady, P.; Hannick, L.; Hariharan, P.; Hilsenbeck, S.; Hindenach, B.; Hoadley, K. A.; Hong, R.; Hostetter, G.; Hsieh, J. J.; Hu, Y.; Ittmann, M. M.; Jaehnig, E.; Jewell, S. D.; Ji, J.; Jones, C. D.; Karabon, R.; Ketchum, K. A.; Khan, M.; Kim, B.-J.; Krek, A.; Krubit, T.; Kumar-Sinha, C.; Leprevost, F. D.; Lewis, M.; Li, Q. K.; Li, Y.; Liu, H.; Lubinski, J.; Ma, W.; Madan, R.; Malc, E.; Malovannaya, A.; Mareedu, S.; Markey, S. P.; Marrero-Oliveras, A.; Martignetti, J.; McDermott, J.; McGarvey, P. B.; McGee, J.; Mieczkowski, P.; Modugno, F.; Montgomery, R.; Newton, C. J.; Omenn, G. S.; Paulovich, A. G.; Perou, A. M.; Petralia, F.; Piehowski, P.; Polonskaya, L.; Qi, L.; Richey, S.; Robinson, K.; Roche, N.; Rohrer, D. C.; Schadt, E. E.; Schnaubelt, M.; Shi, Y.; Skelly, T.; Sokoll, L. J.; Song, X.; Stein, S. E.; Suh, J.; Tan, D.; Tansil, D.; Teo, G. C.; Thangudu, R. R.; Tognon, C.; Traer, E.; Tyner, J.; Um, K. S.; Valley, D. R.; Vatanian, N.; Vats, P.; Velvulou, U.; Vernon, M.; Wang, L.-B.; Wang, Y.; Webster, A.; Westbrook, T.; Wheeler, D.;

Whiteaker, J. R.; Wilson, G. D.; Zakhartsev, Y.; Zelt, R.; Zhang, H.; Zhang, Y.; Zhang, Z.; Zhao, G. Proteogenomic Characterization of Endometrial Carcinoma. *Cell* **2020**, *180* (4), 729-748.e26. https://doi.org/10.1016/j.cell.2020.01.026.

(20) Vasaikar, S.; Huang, C.; Wang, X.; Petyuk, V. A.; Savage, S. R.; Wen, B.; Dou, Y.; Zhang, Y.; Shi, Z.; Arshad, O. A.; Gritsenko, M. A.; Zimmerman, L. J.; McDermott, J. E.; Clauss, T. R.; Moore, R. J.; Zhao, R.; Monroe, M. E.; Wang, Y.-T.; Chambers, M. C.; Slebos, R. J. C.; Lau, K. S.; Mo, Q.; Ding, L.; Ellis, M.; Thiagarajan, M.; Kinsinger, C. R.; Rodriguez, H.; Smith, R. D.; Rodland, K. D.; Liebler, D. C.; Liu, T.; Zhang, B.; Clinical Proteomic Tumor Analysis Consortium. Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. *Cell* **2019**, *177* (4), 1035-1049.e19. https://doi.org/10.1016/j.cell.2019.03.030.

(21) Gillette, M. A.; Satpathy, S.; Cao, S.; Dhanasekaran, S. M.; Vasaikar, S. V.; Krug, K.; Petralia, F.; Li, Y.; Liang, W.-W.; Reva, B.; Krek, A.; Ji, J.; Song, X.; Liu, W.; Hong, R.; Yao, L.; Blumenberg, L.; Savage, S. R.; Wendl, M. C.; Wen, B.; Li, K.; Tang, L. C.; MacMullan, M. A.; Avanessian, S. C.; Kane, M. H.; Newton, C. J.; Cornwell, M.; Kothadia, R. B.; Ma, W.; Yoo, S.; Mannan, R.; Vats, P.; Kumar-Sinha, C.; Kawaler, E. A.; Omelchenko, T.; Colaprico, A.; Geffen, Y.; Maruvka, Y. E.; da Veiga Leprevost, F.; Wiznerowicz, M.; Gümüş, Z. H.; Veluswamy, R. R.; Hostetter, G.; Heiman, D. I.; Wyczalkowski, M. A.; Hiltke, T.; Mesri, M.; Kinsinger, C. R.; Boja, E. S.; Omenn, G. S.; Chinnaiyan, A. M.; Rodriguez, H.; Li, Q. K.; Jewell, S. D.; Thiagarajan, M.; Getz, G.; Zhang, B.; Fenyö, D.; Ruggles, K. V.; Cieslik, M. P.; Robles, A. I.; Clauser, K. R.; Govindan, R.; Wang, P.; Nesvizhskii, A. I.; Ding, L.; Mani, D. R.; Carr, S. A.; Clinical Proteomic Tumor Analysis Consortium. Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma. *Cell* **2020**, *182* (1), 200-225.e35. https://doi.org/10.1016/j.cell.2020.06.013.

(22) Clark, D. J.; Dhanasekaran, S. M.; Petralia, F.; Pan, J.; Song, X.; Hu, Y.; da Veiga Leprevost, F.; Reva, B.; Lih, T.-S. M.; Chang, H.-Y.; Ma, W.; Huang, C.; Ricketts, C. J.; Chen, L.; Krek, A.; Li, Y.; Rykunov, D.; Li, Q. K.; Chen, L. S.; Ozbek, U.; Vasaikar, S.; Wu, Y.; Yoo, S.; Chowdhury, S.; Wyczalkowski, M. A.; Ji, J.; Schnaubelt, M.; Kong, A.; Sethuraman, S.; Avtonomov, D. M.; Ao, M.; Colaprico, A.; Cao, S.; Cho, K.-C.; Kalayci, S.; Ma, S.; Liu, W.; Ruggles, K.; Calinawan, A.; Gümüş, Z. H.; Geiszler, D.; Kawaler, E.; Teo, G. C.; Wen, B.; Zhang, Y.; Keegan, S.; Li, K.; Chen, F.; Edwards, N.; Pierorazio, P. M.; Chen, X. S.; Pavlovich, C. P.; Hakimi, A. A.; Brominski, G.; Hsieh, J. J.; Antczak, A.; Omelchenko, T.; Lubinski, J.; Wiznerowicz, M.; Linehan, W. M.; Kinsinger, C. R.; Thiagarajan, M.; Boja, E. S.; Mesri, M.; Hiltke, T.; Robles, A. I.; Rodriguez, H.; Qian, J.; Fenyö, D.; Zhang, B.; Ding, L.; Schadt, E.; Chinnaiyan, A. M.; Zhang, Z.; Omenn, G. S.; Cieslik, M.; Chan, D. W.; Nesvizhskii, A. I.; Wang, P.; Zhang, H.; Clinical Proteomic Tumor Analysis Consortium. Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma. *Cell* **2019**, *179* (4), 964-983.e31. https://doi.org/10.1016/j.cell.2019.10.007.

(23) Wang, L.-B.; Karpova, A.; Gritsenko, M. A.; Kyle, J. E.; Cao, S.; Li, Y.; Rykunov, D.; Colaprico, A.; Rothstein, J. H.; Hong, R.; Stathias, V.; Cornwell, M.; Petralia, F.; Wu, Y.; Reva, B.; Krug, K.; Pugliese, P.; Kawaler, E.; Olsen, L. K.; Liang, W.-W.; Song, X.; Dou, Y.; Wendl, M. C.; Caravan, W.; Liu, W.; Cui Zhou, D.; Ji, J.; Tsai, C.-F.; Petyuk, V. A.; Moon, J.; Ma, W.; Chu, R. K.; Weitz, K. K.; Moore, R. J.; Monroe, M. E.; Zhao, R.; Yang, X.; Yoo, S.; Krek, A.; Demopoulos, A.; Zhu, H.; Wyczalkowski, M. A.; McMichael, J. F.; Henderson, B. L.; Lindgren, C. M.; Boekweg, H.; Lu, S.; Baral, J.; Yao, L.; Stratton, K. G.; Bramer, L. M.; Zink, E.; Couvillion, S. P.; Bloodsworth, K. J.; Satpathy, S.; Sieh, W.; Boca, S. M.; Schürer, S.; Chen, F.; Wiznerowicz, M.; Ketchum, K. A.; Boja, E. S.; Kinsinger, C. R.; Robles, A. I.; Hiltke, T.; Thiagarajan, M.; Nesvizhskii, A. I.; Zhang, B.; Mani, D. R.; Ceccarelli, M.; Chen, X. S.;

Cottingham, S. L.; Li, Q. K.; Kim, A. H.; Fenyö, D.; Ruggles, K. V.; Rodriguez, H.; Mesri, M.; Payne, S. H.; Resnick, A. C.; Wang, P.; Smith, R. D.; Iavarone, A.; Chheda, M. G.; Barnholtz-Sloan, J. S.; Rodland, K. D.; Liu, T.; Ding, L.; Clinical Proteomic Tumor Analysis Consortium. Proteogenomic and Metabolomic Characterization of Human Glioblastoma. *Cancer Cell* **2021**, *39* (4), 509-528.e20. https://doi.org/10.1016/j.ccell.2021.01.006.

(24) Satpathy, S.; Krug, K.; Jean Beltran, P. M.; Savage, S. R.; Petralia, F.; Kumar-Sinha, C.; Dou, Y.; Reva, B.; Kane, M. H.; Avanessian, S. C.; Vasaikar, S. V.; Krek, A.; Lei, J. T.; Jaehnig, E. J.; Omelchenko, T.; Geffen, Y.; Bergstrom, E. J.; Stathias, V.; Christianson, K. E.; Heiman, D. I.; Cieslik, M. P.; Cao, S.; Song, X.; Ji, J.; Liu, W.; Li, K.; Wen, B.; Li, Y.; Gümüş, Z. H.; Selvan, M. E.; Soundararajan, R.; Visal, T. H.; Raso, M. G.; Parra, E. R.; Babur, Ö.; Vats, P.; Anand, S.; Schraink, T.; Cornwell, M.; Rodrigues, F. M.; Zhu, H.; Mo, C.-K.; Zhang, Y.; da Veiga Leprevost, F.; Huang, C.; Chinnaiyan, A. M.; Wyczalkowski, M. A.; Omenn, G. S.; Newton, C. J.; Schurer, S.; Ruggles, K. V.; Fenyö, D.; Jewell, S. D.; Thiagarajan, M.; Mesri, M.; Rodriguez, H.; Mani, S. A.; Udeshi, N. D.; Getz, G.; Suh, J.; Li, Q. K.; Hostetter, G.; Paik, P. K.; Dhanasekaran, S. M.; Govindan, R.; Ding, L.; Robles, A. I.; Clauser, K. R.; Nesvizhskii, A. I.; Wang, P.; Carr, S. A.; Zhang, B.; Mani, D. R.; Gillette, M. A.; Clinical Proteomic Tumor Analysis Consortium. A Proteogenomic Portrait of Lung Squamous Cell Carcinoma. *Cell* **2021**, *184* (16), 4348-4371.e40. https://doi.org/10.1016/j.cell.2021.07.016.

(25) Lindgren, C. M.; Adams, D. W.; Kimball, B.; Boekweg, H.; Tayler, S.; Pugh, S. L.; Payne, S. H. Simplified and Unified Access to Cancer Proteogenomic Data. *J. Proteome Res.* **2021**, *20* (4), 1902–1910. https://doi.org/10.1021/acs.jproteome.0c00919.

(26) Giurgiu, M.; Reinhard, J.; Brauner, B.; Dunger-Kaltenbach, I.; Fobo, G.; Frishman, G.; Montrone, C.; Ruepp, A. CORUM: The Comprehensive Resource of Mammalian Protein Complexes-2019. *Nucleic Acids Res.* **2019**, *47* (D1), D559–D563. https://doi.org/10.1093/nar/gky973.

(27) Szklarczyk, D.; Gable, A. L.; Nastou, K. C.; Lyon, D.; Kirsch, R.; Pyysalo, S.; Doncheva, N. T.; Legeay, M.; Fang, T.; Bork, P.; Jensen, L. J.; von Mering, C. The STRING Database in 2021: Customizable Protein-Protein Networks, and Functional Characterization of User-Uploaded Gene/Measurement Sets. *Nucleic Acids Res.* **2021**, *49* (D1), D605–D612. https://doi.org/10.1093/nar/gkaa1074.

(28) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12* (85), 2825–2830.

(29) Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc., 2017; Vol. 30.

(30) Kursa, M. B.; Rudnicki, W. R. Feature Selection with the Boruta Package. *J. Stat. Softw.* **2010**, *36* (11). https://doi.org/10.18637/jss.v036.i11.

(31) Hagberg, A.; Swart, P.; S Chult, D. *Exploring Network Structure, Dynamics, and Function Using Networkx*; LA-UR-08-05495; LA-UR-08-5495; Los Alamos National Lab. (LANL), Los Alamos, NM (United States), 2008.

(32) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13* (11), 2498–2504. https://doi.org/10.1101/gr.1239303.

(33) Zheng, F.; Zhang, S.; Churas, C.; Pratt, D.; Bahar, I.; Ideker, T. HiDeF: Identifying Persistent

Structures in Multiscale ‘omics Data. *Genome Biol.* **2021**, *22* (1), 21. https://doi.org/10.1186/s13059-020-02228-4.

(34) Singhal, A.; Cao, S.; Churas, C.; Pratt, D.; Fortunato, S.; Zheng, F.; Ideker, T. Multiscale Community Detection in Cytoscape. *PLOS Comput. Biol.* **2020**, *16* (10), e1008239. https://doi.org/10.1371/journal.pcbi.1008239.

(35) Traag, V. A.; Waltman, L.; van Eck, N. J. From Louvain to Leiden: Guaranteeing Well-Connected Communities. *Sci. Rep.* **2019**, *9* (1), 5233. https://doi.org/10.1038/s41598-019-41695-z.

(36) Reimand, J.; Arak, T.; Adler, P.; Kolberg, L.; Reisberg, S.; Peterson, H.; Vilo, J. G:Profiler—a Web Server for Functional Interpretation of Gene Lists (2016 Update). *Nucleic Acids Res.* **2016**, *44* (W1), W83–W89. https://doi.org/10.1093/nar/gkw199.

(37) Tang, Y.; Li, M.; Wang, J.; Pan, Y.; Wu, F.-X. CytoNCA: A Cytoscape Plugin for Centrality Analysis and Evaluation of Protein Interaction Networks. *Biosystems* **2015**, *127*, 67–72. https://doi.org/10.1016/j.biosystems.2014.11.005.

(38) Doncheva, N. T.; Morris, J. H.; Gorodkin, J.; Jensen, L. J. Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data. *J. Proteome Res.* **2019**, *18* (2), 623–632. https://doi.org/10.1021/acs.jproteome.8b00702.

(39) Yu, G.; He, Q.-Y. ReactomePA: An R/Bioconductor Package for Reactome Pathway Analysis and Visualization. *Mol. Biosyst.* **2016**, *12* (2), 477–479. https://doi.org/10.1039/C5MB00663E.

(40) Wu, T.; Hu, E.; Xu, S.; Chen, M.; Guo, P.; Dai, Z.; Feng, T.; Zhou, L.; Tang, W.; Zhan, L.; Fu, X.; Liu, S.; Bo, X.; Yu, G. ClusterProfiler 4.0: A Universal Enrichment Tool for Interpreting Omics Data. *The Innovation* **2021**, *2* (3), 100141. https://doi.org/10.1016/j.xinn.2021.100141.

(41) Gu, Z.; Gu, L.; Eils, R.; Schlesner, M.; Brors, B. Circlize Implements and Enhances Circular Visualization in R. *Bioinformatics* **2014**, *30* (19), 2811–2812. https://doi.org/10.1093/bioinformatics/btu393.

(42) Waskom, M. Seaborn: Statistical Data Visualization. *J. Open Source Softw.* **2021**, *6* (60), 3021. https://doi.org/10.21105/joss.03021.

(43) Mick, D. U.; Dennerlein, S.; Wiese, H.; Reinhold, R.; Pacheu-Grau, D.; Lorenzi, I.; Sasarman, F.; Weraarpachai, W.; Shoubridge, E. A.; Warscheid, B.; Rehling, P. MITRAC Links Mitochondrial Protein Translocation to Respiratory-Chain Assembly and Translational Regulation. *Cell* **2012**, *151* (7), 1528–1541. https://doi.org/10.1016/j.cell.2012.11.053.

(44) Signes, A.; Fernandez-Vizarra, E. Assembly of Mammalian Oxidative Phosphorylation Complexes I–V and Supercomplexes. *Essays Biochem.* **2018**, *62* (3), 255–270. https://doi.org/10.1042/EBC20170098.

(45) Hirano, Y.; Hendil, K. B.; Yashiroda, H.; Iemura, S.; Nagane, R.; Hioki, Y.; Natsume, T.; Tanaka, K.; Murata, S. A Heterodimeric Complex That Promotes the Assembly of Mammalian 20S Proteasomes. *Nature* **2005**, *437* (7063), 1381–1385. https://doi.org/10.1038/nature04106.

(46) The GTEx Consortium. The GTEx Consortium Atlas of Genetic Regulatory Effects across Human Tissues. *Science* **2020**, *369* (6509), 1318–1330. https://doi.org/10.1126/science.aaz1776.

(47) Sasaki, K.; Hamazaki, J.; Koike, M.; Hirano, Y.; Komatsu, M.; Uchiyama, Y.; Tanaka, K.; Murata, S. PAC1 Gene Knockout Reveals an Essential Role of Chaperone-Mediated 20S Proteasome Biogenesis and Latent 20S Proteasomes in Cellular Homeostasis. *Mol. Cell. Biol.* **2010**, *30* (15), 3864–3874. https://doi.org/10.1128/MCB.00216-10.

(48) Tang, D.; Sheng, J.; Xu, L.; Zhan, X.; Liu, J.; Jiang, H.; Shu, X.; Liu, X.; Zhang, T.; Jiang, L.; Zhou, C.; Li, W.; Cheng, W.; Li, Z.; Wang, K.; Lu, K.; Yan, C.; Qi, S. Cryo-EM Structure of C9ORF72–SMCR8–WDR41 Reveals the Role as a GAP for Rab8a and Rab11a. *Proc. Natl.*

Acad. Sci. **2020**, *117* (18), 9876–9883. https://doi.org/10.1073/pnas.2002110117.

(49) Sellier, C.; Campanari, M.-L.; Julie Corbier, C.; Gaucherot, A.; Kolb-Cheynel, I.; Oulad-Abdelghani, M.; Ruffenach, F.; Page, A.; Ciura, S.; Kabashi, E.; Charlet-Berguerand, N. Loss of C9ORF72 Impairs Autophagy and Synergizes with PolyQ Ataxin-2 to Induce Motor Neuron Dysfunction and Cell Death. *EMBO J.* **2016**, *35* (12), 1276–1297. https://doi.org/10.15252/embj.201593350.

(50) Sullivan, P. M.; Zhou, X.; Robins, A. M.; Paushter, D. H.; Kim, D.; Smolka, M. B.; Hu, F. The ALS/FTLD Associated Protein C9orf72 Associates with SMCR8 and WDR41 to Regulate the Autophagy-Lysosome Pathway. *Acta Neuropathol. Commun.* **2016**, *4* (1), 51. https://doi.org/10.1186/s40478-016-0324-5.

(51) Kamer, K. J.; Mootha, V. K. MICU1 and MICU2 Play Nonredundant Roles in the Regulation of the Mitochondrial Calcium Uniporter. *EMBO Rep.* **2014**, *15* (3), 299–307. https://doi.org/10.1002/embr.201337946.

(52) Plovanich, M.; Bogorad, R. L.; Sancak, Y.; Kamer, K. J.; Strittmatter, L.; Li, A. A.; Girgis, H. S.; Kuchimanchi, S.; De Groot, J.; Speciner, L.; Taneja, N.; Oshea, J.; Koteliansky, V.; Mootha, V. K. MICU2, a Paralog of MICU1, Resides within the Mitochondrial Uniporter Complex to Regulate Calcium Handling. *PloS One* **2013**, *8* (2), e55785. https://doi.org/10.1371/journal.pone.0055785.

(53) Hoffman, A.; Taleski, G.; Sontag, E. The Protein Serine/Threonine Phosphatases PP2A, PP1 and Calcineurin: A Triple Threat in the Regulation of the Neuronal Cytoskeleton. *Mol. Cell. Neurosci.* **2017**, *84*, 119–131. https://doi.org/10.1016/j.mcn.2017.01.005.

(54) Rusnak, F.; Mertz, P. Calcineurin: Form and Function. *Physiol. Rev.* **2000**, *80* (4), 1483–1521. https://doi.org/10.1152/physrev.2000.80.4.1483.

(55) Liberzon, A.; Birger, C.; Thorvaldsdóttir, H.; Ghandi, M.; Mesirov, J. P.; Tamayo, P. The Molecular Signatures Database (MSigDB) Hallmark Gene Set Collection. *Cell Syst.* **2015**, *1* (6), 417–425. https://doi.org/10.1016/j.cels.2015.12.004.

(56) Lau, E.; Venkatraman, V.; Thomas, C. T.; Wu, J. C.; Van Eyk, J. E.; Lam, M. P. Y. Identifying High-Priority Proteins Across the Human Diseasome Using Semantic Similarity. *J. Proteome Res.* **2018**, *17* (12), 4267–4278. https://doi.org/10.1021/acs.jproteome.8b00393.

(57) Lam, M. P. Y.; Venkatraman, V.; Xing, Y.; Lau, E.; Cao, Q.; Ng, D. C. M.; Su, A. I.; Ge, J.; Van Eyk, J. E.; Ping, P. Data-Driven Approach To Determine Popular Proteins for Targeted Proteomics Translation of Six Organ Systems. *J. Proteome Res.* **2016**, *15* (11), 4126–4134. https://doi.org/10.1021/acs.jproteome.6b00095.

(58) Koc, E. C.; Burkhart, W.; Blackburn, K.; Moyer, M. B.; Schlatzer, D. M.; Moseley, A.; Spremulli, L. L. The Large Subunit of the Mammalian Mitochondrial Ribosome: ANALYSIS OF THE COMPLEMENT OF RIBOSOMAL PROTEINS PRESENT*. *J. Biol. Chem.* **2001**, *276* (47), 43958–43969. https://doi.org/10.1074/jbc.M106510200.

(59) Polianskyte, Z.; Peitsaro, N.; Dapkunas, A.; Liobikas, J.; Soliymani, R.; Lalowski, M.; Speer, O.; Seitsonen, J.; Butcher, S.; Cereghetti, G. M.; Linder, M. D.; Merckel, M.; Thompson, J.; Eriksson, O. LACTB Is a Filament-Forming Protein Localized in Mitochondria. *Proc. Natl. Acad. Sci.* **2009**, *106* (45), 18960–18965. https://doi.org/10.1073/pnas.0906734106.

(60) Keckesova, Z.; Donaher, J. L.; De Cock, J.; Freinkman, E.; Lingrell, S.; Bachovchin, D. A.; Bierie, B.; Tischler, V.; Noske, A.; Okondo, M. C.; Reinhardt, F.; Thiru, P.; Golub, T. R.; Vance, J. E.; Weinberg, R. A. LACTB Is a Tumour Suppressor That Modulates Lipid Metabolism and Cell State. *Nature* **2017**, *543* (7647), 681–686. https://doi.org/10.1038/nature21408.

(61) Kim, T.; Tao-Cheng, J.-H.; Eiden, L. E.; Loh, Y. P. Chromogranin A, an “On/Off” Switch Controlling Dense-Core Secretory Granule Biogenesis. *Cell* **2001**, *106* (4), 499–509.

https://doi.org/10.1016/S0092-8674(01)00459-7.

(62) Aakre, C. D.; Herrou, J.; Phung, T. N.; Perchuk, B. S.; Crosson, S.; Laub, M. T. Evolving New Protein-Protein Interaction Specificity through Promiscuous Intermediates. *Cell* **2015**, *163* (3), 594–606. https://doi.org/10.1016/j.cell.2015.09.055.

(63) Payne, S. H. The Utility of Protein and MRNA Correlation. *Trends Biochem. Sci.* **2015**, *40* (1), 1–3. https://doi.org/10.1016/j.tibs.2014.10.010.

(64) Sanchez-Vega, F.; Mina, M.; Armenia, J.; Chatila, W. K.; Luna, A.; La, K. C.; Dimitriadoy, S.; Liu, D. L.; Kantheti, H. S.; Saghafinia, S.; Chakravarty, D.; Daian, F.; Gao, Q.; Bailey, M. H.; Liang, W.-W.; Foltz, S. M.; Shmulevich, I.; Ding, L.; Heins, Z.; Ochoa, A.; Gross, B.; Gao, J.; Zhang, H.; Kundra, R.; Kandoth, C.; Bahceci, I.; Dervishi, L.; Dogrusoz, U.; Zhou, W.; Shen, H.; Laird, P. W.; Way, G. P.; Greene, C. S.; Liang, H.; Xiao, Y.; Wang, C.; Iavarone, A.; Berger, A. H.; Bivona, T. G.; Lazar, A. J.; Hammer, G. D.; Giordano, T.; Kwong, L. N.; McArthur, G.; Huang, C.; Tward, A. D.; Frederick, M. J.; McCormick, F.; Meyerson, M.; Cancer Genome Atlas Research Network; Van Allen, E. M.; Cherniack, A. D.; Ciriello, G.; Sander, C.; Schultz, N. Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* **2018**, *173* (2), 321-337.e10. https://doi.org/10.1016/j.cell.2018.03.035.