

# Multi-Dataset Integration and Residual Connections Improve Proteome Prediction from Transcriptomics Using Deep Learning

This manuscript ([permalink](#)) was automatically generated from [xomicsdatascience/transcriptome-proteome-nas-manubot@5a003b6](#) on July 8, 2024.

## Authors

---

- **Caleb W. Cranney**

 [0000-0001-8482-758X](#) ·  [CCranney](#) ·  [CalebCranney](#)

Department of Computational Biomedicine, Cedars Sinai Medical Center · Funded by Grant R35GM142502, National Institute of General Medical Sciences (NIGMS)

- **Jesse G. Meyer**

 [0000-0003-2753-3926](#) ·  [jessegmeyerlab](#) ·  [j\\_my\\_sci](#)

Department of Computational Biomedicine, Cedars Sinai Medical Center

✉ — Correspondence possible via [GitHub Issues](#)

# Abstract

---

Proteomes are well known to poorly correlate with transcriptomes measured from the same sample. While connected, the complex processes that impact the relationships between transcript and protein quantities remains an open research topic. Many studies have attempted to predict proteomes from transcriptomes with limited success. Here we use publicly available data from the Clinical Proteomics Tumor Analysis Consortium to show that deep learning models designed by neural architecture search (NAS) achieve improved prediction accuracy of proteome quantities from transcriptomics. We find that this benefit is largely due to including a residual connection in the architecture that allows input information to be remembered near the end of the network. Finally, we explore which groups of transcripts are functionally important for protein prediction using model interpretation with SHAP.

# Introduction

---

The central dogma of biology posits a linear flow of information from genetic encoding to mRNA transcripts to functional proteins. However, this seemingly straightforward relationship belies a more intricate reality, where the interactions between omic layers are multifaceted and complex. Elucidating these relationships is crucial for understanding biological systems in both healthy and diseased states. By disentangling these interactions, researchers can identify markers and patterns specific to complex diseases, ultimately enabling the development of targeted treatments.

Among the multi-omic data type relationships, the connection between transcripts and their corresponding proteins is particularly enigmatic. Despite being quantifiable and having a direct derivative relationship, their relative quantities often exhibit only weak correlations [1], [2]. Research has identified several possible causes for this discrepancy, including alternate rates of protein generation and decay [3], [4], [5], varying reactions to environmental stimuli [6], [7], or simply systematic experimentation error bias [8], [9]. Recent works have even found that the most predictive transcripts of a protein include those that are involved in protein-protein interactions [10]. While predicting protein quantity through direct transcript-to-protein correlation remains elusive, using contextual transcript information may reveal more complex proteomic-transcriptomic relationships that could be leveraged to predict one from the other. Moreover, predicting proteins from transcripts could reduce the time and financial costs associated with future studies, as transcript quantification is generally easier to accomplish than protein quantification [11].

The National Cancer Institute's (NCI) Clinical Proteomic Tumor Analysis Consortium ([CPTAC](#)) provides a valuable resource for exploring this issue, offering multi-omic datasets that enable research into healthy and cancerous disease states. Specifically, CPTAC datasets comprise transcriptomic and proteomic data (among others) from various cancer types and adjacent healthy tissue, which makes this data useful for detection of inter-omic relationships. Several research collaborations have occurred to use this data to study RNA-protein relationships. The 2017 [NCI-CPTAC DREAM Proteogenomics challenge](#), a sub challenge of which aimed to predict protein quantities from transcript abundance, is a notable example. Contestants used a variety of methods, including random forest regression, genetic models, spline regression, linear regression, and elastic net methods, as well as ensemble combinations [12]. Using a test set only from breast cancer and ovarian cancer, the winning model achieved a pearson correlation of 0.41 and 0.47 between the true and predicted protein quantities, respectively [13]. Notably, deep learning neural networks were not strongly represented in the challenge results.

Since the close of the challenge, research into deep learning has expanded significantly, with a focus on replicating human behaviors like computer vision and natural language processing (NLP). However, the underlying principles of deep learning are equally applicable to biological data [14], [15], [16]. The key challenge lies in developing a model architecture that best fits a given problem. One underutilized strategy for deep learning model design is the NAS [17]. NAS serves a function like hyperparameter tuning, in that a range of values for specific hyperparameters are evaluated to obtain the optimal configurations. In the case of NAS, the concept is expanded to include model architecture in addition to hyperparameters, designing optimal and unique model architectures in an automated fashion. While NAS has been applied to genomic data [18], its application to predicting proteomic data from transcriptomic data remains unexplored.

We previously showed how machine learning can accurately predict the metabolome from the proteome, and how model interpretation revealed important biological insights [19]. Here, we extend that work to transcript-to-protein deep learning prediction models and demonstrate that utilizing NAS improved the accuracy. Furthermore, we highlight the potential of model interpretation to identify

patterns in transcript-protein relations that underpin biological processes characteristic of a disease state.

# Methods

---

## Data Acquisition and Preprocessing

To facilitate readability and reproducibility, the code for downloading, processing, and splitting data was developed as a multi-class in-house package. The bridge design pattern specifically was used to allow for interchangeability of data source input to expand beyond CPTAC in the future, as well as for allowing custom processing and splitting depending on the experiment. This involved writing an abstract parent data processing class, and individual child classes would utilize compartmentalized class components specific to the experiment. This was done to enable external researchers to trace data processing workflows with ease.

CPTAC data was downloaded directly from zenodo using the cptac python package [\[20\]](#) separately for each cancer type. Analyses were only performed using transcripts and proteins common between all datasets.

Cancer-specific datasets were normalized independently of one another. For experiments where each dataset required an identical train-validation split, this normalization was calculated on the training partition then applied to the training and validation partitions both. For the five by two cross validation experiments, it was determined that dataset-specific analyses would be enhanced by the inclusion of all other datasets in the training dataset [Supplementary Figure 1]. In these instances, the train-validation split was applied only to the target dataset, with a split of 0.45, 0.45, and 0.1 for training, validation and testing, respectively. All non-target datasets were normalized on the entire dataset as the training partition, ignoring validation or testing partitioning entirely. In a standard data partition, each dataset had a split of 0.8, 0.1 and 0.1 for training, validation and testing, respectively.

The final dataset contained data for breast cancer (BRCA), kidney cancer (CCRCC), colon cancer (COAD), brain cancer (GBM), squamous cell cancer (HNSCC), lung cancer (LSCC and LUAD), ovarian cancer (OV), and pancreatic cancer (PDAC) as well as adjacent healthy tissue. 59286 transcripts and 7822 proteins were shared across all datasets with a combined number of 1256 samples.

## References

---

1. **Correlation between Protein and mRNA Abundance in Yeast**  
Steven P Gygi, Yvan Rochon, BRobert Franza, Ruedi Aebersold  
*Molecular and Cellular Biology* (1999-03-01) <https://doi.org/grtw7n>  
DOI: [10.1128/mcb.19.3.1720](https://doi.org/10.1128/mcb.19.3.1720) · PMID: [10022859](https://pubmed.ncbi.nlm.nih.gov/10022859/) · PMCID: [PMC83965](https://pubmed.ncbi.nlm.nih.gov/PMC83965/)
2. **The utility of protein and mRNA correlation**  
Samuel H Payne  
*Trends in Biochemical Sciences* (2015-01) <https://doi.org/gf2w7q>  
DOI: [10.1016/j.tibs.2014.10.010](https://doi.org/10.1016/j.tibs.2014.10.010) · PMID: [25467744](https://pubmed.ncbi.nlm.nih.gov/25467744/) · PMCID: [PMC4776753](https://pubmed.ncbi.nlm.nih.gov/PMC4776753/)
3. **Insights into the regulation of protein abundance from proteomic and transcriptomic analyses**  
Christine Vogel, Edward M Marcotte  
*Nature Reviews Genetics* (2012-03-13) <https://doi.org/bg9g>  
DOI: [10.1038/nrg3185](https://doi.org/10.1038/nrg3185) · PMID: [22411467](https://pubmed.ncbi.nlm.nih.gov/22411467/) · PMCID: [PMC3654667](https://pubmed.ncbi.nlm.nih.gov/PMC3654667/)
4. **Database for mRNA Half-Life of 19 977 Genes Obtained by DNA Microarray Analysis of Pluripotent and Differentiating Mouse Embryonic Stem Cells**  
LV Sharova, AA Sharov, T Nedorezov, Y Piao, N Shaik, MSH Ko  
*DNA Research* (2009-01-08) <https://doi.org/cvg6wg>  
DOI: [10.1093/dnares/dsn030](https://doi.org/10.1093/dnares/dsn030) · PMID: [19001483](https://pubmed.ncbi.nlm.nih.gov/19001483/) · PMCID: [PMC2644350](https://pubmed.ncbi.nlm.nih.gov/PMC2644350/)
5. **Global Protein Stability Profiling in Mammalian Cells**  
Hsueh-Chi Sherry Yen, Qikai Xu, Danny M Chou, Zhenming Zhao, Stephen J Elledge  
*Science* (2008-11-07) <https://doi.org/dzv5xx>  
DOI: [10.1126/science.1160489](https://doi.org/10.1126/science.1160489) · PMID: [18988847](https://pubmed.ncbi.nlm.nih.gov/18988847/)
6. **On the Dependency of Cellular Protein Levels on mRNA Abundance**  
Yansheng Liu, Andreas Beyer, Ruedi Aebersold  
*Cell* (2016-04) <https://doi.org/f8kc6z>  
DOI: [10.1016/j.cell.2016.03.014](https://doi.org/10.1016/j.cell.2016.03.014) · PMID: [27104977](https://pubmed.ncbi.nlm.nih.gov/27104977/)
7. **Gcn4p, a Master Regulator of Gene Expression, Is Controlled at Multiple Levels by Diverse Signals of Starvation and Stress**  
Alan G Hinnebusch, Krishnamurthy Natarajan  
*Eukaryotic Cell* (2002-02) <https://doi.org/fq6gww>  
DOI: [10.1128/ec.01.1.22-32.2002](https://doi.org/10.1128/ec.01.1.22-32.2002) · PMID: [12455968](https://pubmed.ncbi.nlm.nih.gov/12455968/) · PMCID: [PMC118051](https://pubmed.ncbi.nlm.nih.gov/PMC118051/)
8. **Experimental reproducibility limits the correlation between mRNA and protein abundances in tumor proteomic profiles**  
Swathi Ramachandra Upadhy, Colm J Ryan  
*Cell Reports Methods* (2022-09) <https://doi.org/ggtvfv>  
DOI: [10.1016/j.crmeth.2022.100288](https://doi.org/10.1016/j.crmeth.2022.100288) · PMID: [36160043](https://pubmed.ncbi.nlm.nih.gov/36160043/) · PMCID: [PMC9499981](https://pubmed.ncbi.nlm.nih.gov/PMC9499981/)
9. **Antibody reliability influences observed mRNA–protein correlations in tumour samples**  
Swathi Ramachandra Upadhy, Colm J Ryan  
*Life Science Alliance* (2023-05-11) <https://doi.org/gt3ttb>  
DOI: [10.26508/lsa.202201885](https://doi.org/10.26508/lsa.202201885) · PMID: [37169592](https://pubmed.ncbi.nlm.nih.gov/37169592/) · PMCID: [PMC10176110](https://pubmed.ncbi.nlm.nih.gov/PMC10176110/)
10. **Protein prediction models support widespread post-transcriptional regulation of protein abundance by interacting partners**

Himangi Srivastava, Michael J Lippincott, Jordan Currie, Robert Canfield, Maggie PY Lam, Edward Lau

*PLOS Computational Biology* (2022-11-10) <https://doi.org/grkjgm>

DOI: [10.1371/journal.pcbi.1010702](https://doi.org/10.1371/journal.pcbi.1010702) · PMID: [36356032](https://pubmed.ncbi.nlm.nih.gov/36356032/) · PMCID: [PMC9681107](https://pubmed.ncbi.nlm.nih.gov/PMC9681107/)

11. **Harmonization of quality metrics and power calculation in multi-omic studies**

Sonia Tarazona, Leandro Balzano-Nogueira, David Gómez-Cabrero, Andreas Schmidt, Axel Imhof, Thomas Hankemeier, Jesper Tegnér, Johan A Westerhuis, Ana Conesa

*Nature Communications* (2020-06-18) <https://doi.org/gg2pmg>

DOI: [10.1038/s41467-020-16937-8](https://doi.org/10.1038/s41467-020-16937-8) · PMID: [32555183](https://pubmed.ncbi.nlm.nih.gov/32555183/) · PMCID: [PMC7303201](https://pubmed.ncbi.nlm.nih.gov/PMC7303201/)

12. **Challenges in proteogenomics: a comparison of analysis methods with the case study of the DREAM proteogenomics sub-challenge**

Tara Eicher, Andrew Patt, Esko Kautto, Raghu Machiraju, Ewy Mathé, Yan Zhang

*BMC Bioinformatics* (2019-12) <https://doi.org/gpfz94>

DOI: [10.1186/s12859-019-3253-z](https://doi.org/10.1186/s12859-019-3253-z) · PMID: [31861998](https://pubmed.ncbi.nlm.nih.gov/31861998/) · PMCID: [PMC6923881](https://pubmed.ncbi.nlm.nih.gov/PMC6923881/)

13. **Joint learning improves protein abundance prediction in cancers**

Hongyang Li, Omer Siddiqui, Hongjiu Zhang, Yuanfang Guan

*BMC Biology* (2019-12) <https://doi.org/gr23js>

DOI: [10.1186/s12915-019-0730-9](https://doi.org/10.1186/s12915-019-0730-9) · PMID: [31870366](https://pubmed.ncbi.nlm.nih.gov/31870366/) · PMCID: [PMC6929375](https://pubmed.ncbi.nlm.nih.gov/PMC6929375/)

14. **Ensemble deep learning in bioinformatics**

Yue Cao, Thomas Andrew Geddes, Jean Yee Hwa Yang, Pengyi Yang

*Nature Machine Intelligence* (2020-08-17) <https://doi.org/ghhnj2>

DOI: [10.1038/s42256-020-0217-y](https://doi.org/10.1038/s42256-020-0217-y)

15. **Deep learning in bioinformatics**

Seonwoo Min, Byunghan Lee, Sungroh Yoon

*Briefings in Bioinformatics* (2016-07-29) <https://doi.org/gcgv8v>

DOI: [10.1093/bib/bbw068](https://doi.org/10.1093/bib/bbw068) · PMID: [27473064](https://pubmed.ncbi.nlm.nih.gov/27473064/)

16. **Review of the Applications of Deep Learning in Bioinformatics**

Yongqing Zhang, Jianrong Yan, Siyu Chen, Meiqin Gong, Dongrui Gao, Min Zhu, Wei Gan

*Current Bioinformatics* (2021-01-01) <https://doi.org/gt3ts9>

DOI: [10.2174/1574893615999200711165743](https://doi.org/10.2174/1574893615999200711165743)

17. **A Comprehensive Survey of Neural Architecture Search**

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-yao Huang, Zhihui Li, Xiaojiang Chen, Xin Wang

*ACM Computing Surveys* (2021-05-24) <https://doi.org/gk76qj>

DOI: [10.1145/3447582](https://doi.org/10.1145/3447582)

18. **An automated framework for efficiently designing deep convolutional neural networks in genomics**

Zijun Zhang, Christopher Y Park, Chandra L Theesfeld, Olga G Troyanskaya

*Nature Machine Intelligence* (2021-03-15) <https://doi.org/gnft7>

DOI: [10.1038/s42256-021-00316-z](https://doi.org/10.1038/s42256-021-00316-z)

19. **Multi-omic integration by machine learning (MIMaL)**

Quinn Dickinson, Andreas Kohler, Martin Ott, Jesse G Meyer

*Bioinformatics* (2022-09-15) <https://doi.org/gg93d7>

DOI: [10.1093/bioinformatics/btac631](https://doi.org/10.1093/bioinformatics/btac631) · PMID: [36106996](https://pubmed.ncbi.nlm.nih.gov/36106996/) · PMCID: [PMC9801967](https://pubmed.ncbi.nlm.nih.gov/PMC9801967/)

20. **Simplified and Unified Access to Cancer Proteogenomic Data**

Caleb M Lindgren, David W Adams, Benjamin Kimball, Hannah Boekweg, Sadie Tayler, Samuel L Pugh, Samuel H Payne

