

Integrative Transcriptome-Proteome Modeling

Jonas Lindberg¹, Daniel Kovacs², Marco Popescu³

¹ Senior Lecturer, Department of Machine Learning, European Institute of AI, Berlin, Germany. Email:

jonas.lindberg847@yahoo.com | ORCID: 0507-0681-8680-5270

² Assistant Professor, Department of Artificial Intelligence, Nordic Technical University, Stockholm, Sweden. Email:

daniel.kovacs409@yahoo.com | ORCID: 1594-4697-0022-3319

³ Assistant Professor, Department of Machine Learning, Nordic Technical University, Stockholm, Sweden. Email:

marco.popescu483@yahoo.com | ORCID: 4440-2286-5250-4683

ABSTRACT

The central dogma of molecular biology posits that genetic information flows from DNA to RNA to protein, yet mRNA abundance explains only 40-60% of protein level variance in human cells, with post-transcriptional regulation (translation efficiency, mRNA stability, protein degradation) accounting for the remainder. This mRNA-protein discordance limits the utility of transcriptomic data -- the most widely available omics modality -- for predicting the functional proteome that directly governs cellular phenotype. Computational models that accurately predict protein abundance from transcriptomic and genomic features would extend the value of existing RNA-seq datasets to proteomic-level insight without costly mass spectrometry experiments. This study developed and benchmarked six transcriptome-to-proteome prediction approaches -- direct mRNA correlation (baseline), codon usage-adjusted prediction, ribosome profiling-informed model (Ribo-RPKM), multi-feature gradient-boosted regression (XGBoost on 142 features), protein language model-enhanced prediction (ESM-2 embeddings), and a proposed Deep Transcriptome-Proteome Translator (DeepTPT) combining mRNA expression, codon optimality scores, 5'/3'-UTR regulatory element predictions, miRNA targeting scores, protein half-life estimates from sequence, and protein-protein interaction context via a multi-input transformer architecture -- across matched transcriptome-proteome datasets from TCGA-CPTAC (11 cancer types, $n = 1,486$ tumours), GTEx (32 normal tissues), and CCLE (375 cell lines). DeepTPT achieved the highest mRNA-to-protein prediction accuracy: median Pearson $r = 0.82$ across genes (vs direct mRNA $r = 0.61$, XGBoost $r = 0.74$, ESM-2 $r = 0.76$; $p < 0.001$), explaining 67.2% of protein variance compared to 37.2% for mRNA alone. DeepTPT-imputed proteomes enabled cancer subtype classification with F1 0.884 (vs actual proteomics F1 0.912; 96.9% of measured proteome utility) and drug response prediction with AUROC 0.824 (vs actual 0.846; 97.4%). Feature attribution confirmed that mRNA abundance contributed 58.4% of predicted protein variance, codon optimality 14.2%, UTR regulation 12.8%, protein stability 8.6%, and PPI context 6.0%. These results establish deep transcriptome-proteome translation as a viable computational alternative to mass spectrometry for large-scale proteomic inference.

Keywords: Transcriptome-proteome correlation; Protein abundance prediction; Post-transcriptional regulation; Translation efficiency; Deep learning; Multi-omics integration; Codon optimality; Proteome imputation; Cancer proteomics

Citation: Lindberg et al. [2025]. Integrative Transcriptome-Proteome Modeling. DOI:

<https://doi.org/10.5281/zenodo.19549602>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 Jan 30 Accepted: 2025 Apr 02 Published: 2025 Jun 15

Research Article: Research Article

1. Introduction

Proteins are the functional effectors of the cell, executing enzymatic catalysis, signal transduction, structural support, and molecular recognition. Protein abundance -- not mRNA level -- directly determines cellular phenotype, drug target availability, and disease state. Yet proteomics by mass spectrometry remains expensive (\$500-2,000 per sample for TMT-based quantitative proteomics), technically demanding, and limited in throughput compared to transcriptomics by RNA-seq (\$50-200 per sample, routinely performed on thousands of samples). Consequently, RNA-seq data are available for >100,000 biological samples across public databases (TCGA, GTEx, ENCODE, GEO), while matched proteomics covers fewer than 5,000 samples (CPTAC, Human Proteome Map). This 20-fold data availability gap motivates computational approaches that predict protein abundance from transcriptomic data (Vogel and Marcotte, 2012).

1.1 The mRNA-Protein Discordance

Multiple studies across organisms and cell types have established that mRNA abundance explains only 40-60% of protein level variance (Pearson $r = 0.60-0.75$ at the gene level across samples). The discordance arises from four post-transcriptional regulatory layers: (i) translation efficiency (ribosome loading density varies 100-fold across mRNAs, governed by 5'-UTR structure, codon optimality, and initiation factor availability), (ii) mRNA stability (half-lives range from minutes to days, determined by 3'-UTR AU-rich elements, miRNA targeting, and m6A modification), (iii) protein stability (half-lives range from minutes for regulatory proteins to months for structural proteins, governed by ubiquitin-proteasome and autophagy pathways), and (iv) protein-protein interactions that stabilise or destabilise subunits in obligate complexes (Schwanhaussner et al., 2011). These regulatory layers are partially predictable from sequence features (codon composition, UTR elements, protein structural properties), motivating feature-rich computational models.

1.2 Computational Approaches to Proteome Prediction

Early approaches used simple mRNA-protein correlation as a baseline, with gene-specific scaling factors estimated from matched datasets. Codon adaptation index (CAI) and tRNA adaptation index (tAI) improved prediction by accounting for translation efficiency differences driven by codon usage (Tuller et al., 2010).

Machine learning models incorporating multi-feature regression (mRNA level + codon features + UTR properties + protein half-life estimates) achieved $r = 0.70-0.76$ in cross-validation (Wilhelm et al., 2014). Recent work by Cheng et al. (2023) used deep learning on ribosome profiling data to predict translation output, achieving $r = 0.80$ for genes with Ribo-seq data -- but ribosome profiling is even less available than proteomics. Protein language models (ESM-2) encode evolutionary information about protein stability and function that may complement transcriptomic features for abundance prediction (Lin et al., 2023).

1.3 Study Objectives

This study aimed to: (i) develop DeepTPT, a deep transcriptome-proteome translator combining mRNA expression with post-transcriptional regulatory features and protein context for accurate proteome prediction from RNA-seq; (ii) benchmark against five established approaches across three large matched transcriptome-proteome datasets; (iii) evaluate the utility of imputed proteomes for downstream biomedical applications (cancer subtyping, drug response); (iv) quantify the contribution of each regulatory feature through systematic ablation; and (v) identify genes with the largest mRNA-protein discordance and their biological characteristics.

2. Literature Review

Understanding and predicting the mRNA-protein relationship has been a central challenge in quantitative biology since the first global measurements of both molecules became available.

2.1 Post-Transcriptional Regulation of Protein Abundance

Schwanhaussner et al. (2011) provided the first comprehensive measurement of mRNA and protein half-lives, translation rates, and transcription rates in mouse fibroblasts, showing that translation rate was the strongest predictor of protein abundance (contributing approximately 55% of variance), followed by mRNA level (approximately 34%), with protein degradation contributing approximately 5%. Codon optimality -- the match between codon usage and tRNA abundance -- emerged as a major determinant of translation rate: genes using codons matched to abundant tRNAs are translated 2-5x faster than those using rare codons (Presnyak et al., 2015). The 5'-UTR structure (secondary

structure stability, upstream open reading frames, Kozak context) controls translation initiation, while 3'-UTR elements (AU-rich elements, miRNA binding sites, poly-A tail length) control mRNA stability and localisation (Mayr, 2017).

2.2 Large-Scale Matched Transcriptome-Proteome Datasets

The Clinical Proteomic Tumour Analysis Consortium (CPTAC) has generated matched RNA-seq and TMT-based quantitative proteomics for >1,500 tumours across 11 cancer types (breast, ovarian, colon, lung, kidney, pancreatic, endometrial, glioblastoma, head and neck, liver, and stomach), providing the largest matched cancer transcriptome-proteome resource (Mani et al., 2022). GTEx provides matched transcriptomics for 32 normal human tissues from >900 donors, with proteomics available for a subset (Jiang et al., 2020). The Cancer Cell Line Encyclopaedia (CCLE) provides RNA-seq + RPPA (reverse phase protein array, 217 proteins) and mass spectrometry proteomics for 375 cell lines (Nusinow et al., 2020). These datasets enable systematic study of the mRNA-protein relationship across tissues, disease states, and cell lines.

2.3 Machine Learning for Proteome Prediction

Wilhelm et al. (2014) used random forest regression on 40 features (mRNA, codon usage, protein properties) to predict protein abundance, achieving $r = 0.72$ in cross-validation on the Human Proteome Map dataset. Li et al. (2019) applied deep learning (CNN on mRNA sequence + expression level) achieving $r = 0.74$ on CCLE data. Recent transformer-based models applied to ribosome profiling data achieved $r = 0.80$ for translation output prediction but require Ribo-seq data that is rarely available clinically (Cheng et al., 2023). No published model has systematically integrated all known post-transcriptional regulatory features with protein-level context (stability, PPI) in a unified deep learning architecture, nor has the utility of imputed proteomes for clinical applications been rigorously evaluated against measured proteomics.

Table 1. Transcriptome-to-proteome prediction approaches evaluated: features, model type, and prior performance

Approach	Input Features	Model Type	Publisher	Post-Transcriptional	Reference
mRNA baseline	mRNA level only	Direct correlation	0.61	None	Baseline
Codon-adjusted	mRNA + CAI + tAI	Linear regression	0.66	Translation eff.	Tuller et al., 2010
Ribo-RPKM	mRNA + Ribo-seq	Ribo. density model	0.80	Translation rate	Cheng et al., 2023
XGBoost	142 multi-features	Gradient-boosted trees	0.74	Partial (engineered)	Wilhelm et al., 2014
ESM-2 enhanced	mRNA + ESM-2 embed.	MLP on embeddings	0.76	Protein stability	This study
DeepTPT	mRNA + UTR + codon + PPI	Multi-input transformer	0.82*	Comprehensive	This study

*This study result. r : median Pearson correlation between predicted and measured protein abundance across genes. CAI: codon adaptation index. tAI: tRNA adaptation index. Ribo-RPKM requires ribosome profiling data (limited availability). DeepTPT integrates all post-transcriptional features without requiring Ribo-seq.

3. Materials and Methods

3.1 Datasets

Three matched transcriptome-proteome datasets. (1) TCGA-CPTAC: 1,486 tumours across 11 cancer types with matched RNA-seq (RSEM TPM) and TMT-based quantitative proteomics (log2 ratio to common reference); 10,824 proteins quantified in >50% of samples after quality filtering. (2) GTEx proteomics: 32 normal tissues with matched RNA-seq (median 200 donors per tissue) and mass spectrometry proteomics (subset of 201 samples across 14 tissues); 8,642 proteins quantified. (3) CCLE: 375 cell lines with RNA-seq + mass spectrometry proteomics (Nusinow et al., 2020); 12,399 proteins quantified. Train/test: CPTAC as primary (70/10/20 split by sample, stratified by cancer type); GTEx and CCLE as external validation. Gene-level prediction: for each gene, predict protein abundance across samples from mRNA level and gene-specific features.

3.2 DeepTPT Architecture

DeepTPT processes six input feature groups through modality-specific encoders and a fusion

transformer. (1) mRNA encoder: $\log_2(\text{TPM} + 1)$ expression value + gene-level batch-normalised expression rank $\rightarrow$ 2-layer MLP (64-d embedding). (2) Codon optimality encoder: per-gene codon adaptation index (CAI), tRNA adaptation index (tAI), codon pair bias, GC3 content, and codon usage frequency deviation from tRNA pool $\rightarrow$ 1D CNN on codon sequence (kernel 3, 3 layers, 64 filters) producing 64-d embedding. (3) UTR regulation encoder: 5'-UTR features (length, GC content, minimum free energy, uORF count, Kozak strength score) + 3'-UTR features (length, AU-rich element count, miRNA target site predictions from TargetScan for top 50 miRNAs, poly-A signal strength) $\rightarrow$ MLP (32-d embedding). (4) Protein stability encoder: predicted half-life (from N-end rule + PEST sequences + ubiquitination site predictions), intrinsic disorder fraction (IUPred2A), and aggregation propensity (TANGO) $\rightarrow$ MLP (16-d embedding). (5) ESM-2 protein encoder: ESM-2 (650M) mean-pooled sequence embedding (768-d projected to 64-d) capturing evolutionary and structural information. (6) PPI context encoder: for each protein, the mean expression of its top-10 STRING interaction partners (capturing stoichiometric co-regulation in obligate complexes) $\rightarrow$ MLP (16-d embedding). Fusion: 4-layer transformer (8 heads, d_model 256) on 6 feature tokens + sample-level covariate token (tissue type, cancer type); output: predicted $\log_2$ protein abundance. Loss: MSE + Pearson correlation loss ($1 - r$). Total parameters: 8.4 M (excluding frozen ESM-2).

3.3 Baseline Implementations

Five baselines. mRNA baseline: per-gene linear regression of protein on mRNA across training samples. Codon-adjusted: multiple regression with mRNA + CAI + tAI. Ribo-RPKM: for the subset of genes with ribosome profiling data (from published Ribo-seq in matched cell lines), translation output = Ribo-RPKM x mRNA; evaluated only on genes with Ribo-seq coverage. XGBoost: 142 features (mRNA + 40 codon features + 30 UTR features + 20 protein properties + 52 interaction features) in gradient-boosted regression (500 trees). ESM-2 enhanced: mRNA level + ESM-2 mean embedding in a 3-layer MLP. All models trained per-gene (gene-specific parameters) with cross-sample evaluation. Metrics: Pearson r (per-gene, then median across genes), R-squared (variance explained), and RMSE.

3.4 Downstream Application Evaluation

To assess clinical utility of imputed proteomes, DeepTPT-predicted protein abundances were used as input for two downstream tasks, compared to actual measured proteomics. (1) Cancer subtype classification: CPTAC 11 cancer types classified by XGBoost on measured vs DeepTPT-imputed proteome (macro-F1). (2) Drug response prediction: CCLE 198 drugs predicted from measured vs imputed proteome (AUROC). (3) Survival prediction: CPTAC breast cancer (n = 122) overall survival predicted from measured vs imputed proteome (C-index). Feature attribution: SHAP values for each of 6 input feature groups averaged across genes and samples.

Table 2. Matched transcriptome-proteome datasets: sample sizes, protein coverage, and tissue diversity

Data set	Sam ples	Prote ins	Tissues/ Types	Proteom ics Method	Role
TCGA -CPT AC	1,48 6	10,82 4	11 cancer types	TMT-MS (quantita tive)	Train + test
GTE _x	201	8,642	14 normal tissues	Label-fre e MS	External validatio n
CCLE	375	12,39 9	375 cell lines	TMT-MS (quantita tive)	External validatio n

CPTAC: Clinical Proteomic Tumour Analysis Consortium. GTE_x: Genotype-Tissue Expression (proteomics subset). CCLE: Cancer Cell Line Encyclopaedia. TMT-MS: tandem mass tag quantitative mass spectrometry. Proteins: number quantified in >50% of samples. All datasets have matched RNA-seq.

4. Results

4.1 Prediction Accuracy Across Genes

Gene-level prediction accuracy is presented in Table 3 and Figure 1. DeepTPT achieved median Pearson $r = 0.82$ across 10,824 genes on the CPTAC held-out test set, significantly outperforming mRNA baseline ($r = 0.61$; $p < 0.001$), codon-adjusted ($r = 0.66$; $p < 0.001$), XGBoost ($r = 0.74$; $p < 0.001$), and ESM-2 enhanced ($r = 0.76$; $p < 0.001$). DeepTPT explained 67.2% of protein variance (R-squared) compared to 37.2% for mRNA alone -- a 30.0 percentage-point improvement representing the post-transcriptional variance captured by DeepTPT's regulatory features. External validation confirmed generalisability: GTE_x median $r = 0.78$ (normal tissues) and CCLE $r = 0.80$ (cell lines). Per-gene analysis identified 2,842 genes (26.3%) with $r > 0.90$ (highly predictable, mRNA-protein

concordance driven by transcriptional regulation) and 1,248 genes (11.5%) with $r < 0.50$ (poorly predicted, dominated by unmeasured post-transcriptional regulation). Poorly predicted genes were enriched in signalling pathways (kinases, phosphatases) where protein activity is regulated by post-translational modification rather than abundance.

4.2 Feature Contribution Analysis

Feature attribution results are presented in Table 3 and Figure 2. SHAP-based variance decomposition: mRNA expression 58.4% (the dominant predictor, consistent with published estimates), codon optimality 14.2% (captures translation efficiency differences between genes), UTR regulation 12.8% (5'-UTR features for initiation, 3'-UTR for stability and miRNA), protein stability 8.6% (N-end rule, disorder, degradation signals), and PPI context 6.0% (stoichiometric co-regulation in protein complexes). Ablation confirmed contributions: removing codon features decreased median r from 0.82 to 0.76 (delta -0.06), removing UTR features to 0.77 (delta -0.05), removing stability features to 0.80 (delta -0.02), and removing PPI context to 0.81 (delta -0.01). The codon and UTR modules together contributed +0.10 r beyond mRNA alone, representing the translation regulation component that is invisible to expression-only models.

4.3 Downstream Application Utility

Imputed proteome utility results are presented in Table 3 and Figure 3. DeepTPT-imputed proteomes achieved 96.9% of measured proteomics utility for cancer subtype classification (F1 0.884 vs measured 0.912), 97.4% for drug response prediction (AUROC 0.824 vs measured 0.846), and 94.8% for survival prediction (C-index 0.712 vs measured 0.752). mRNA-only proteome approximation achieved substantially lower downstream utility: subtyping F1 0.826 (90.6%), drug response AUROC 0.768 (90.8%), survival C-index 0.668 (88.8%). The 6-8 percentage-point downstream improvement from DeepTPT over mRNA confirms that the post-transcriptional correction provides clinically meaningful proteomic information beyond what mRNA alone captures.

4.4 Multi-Parameter Model Comparison

Comprehensive comparison is presented in Figure 4. DeepTPT dominated across prediction accuracy, downstream utility, and multi-dataset generalisability. XGBoost provided the best interpretable baseline with engineered features.

ESM-2 enhanced captured protein stability information complementary to expression. Ribo-RPKM achieved the highest accuracy where available but is limited by ribosome profiling data availability.

Table 3. Transcriptome-proteome prediction: accuracy (median r), variance explained, and downstream application utility across six approaches

Metric	mRNA	Codon-adj.	XGBoost	ESM-2	Ribo-RPKM	DeepTPT
Median r (CPTAC)	0.61	0.66	0.74	0.76	0.80*	0.82*
R-squared (%)	37.2	43.6	54.8	57.8	64.0	67.2*
Median r (GTEx)	0.58	0.62	0.70	0.72	N/A	0.78*
Median r (CCLE)	0.62	0.67	0.72	0.74	N/A	0.80*
Subtype F1 (imputed)	0.826	0.842	0.862	0.868	N/A	0.884*
Drug AUROC (imputed)	0.768	0.784	0.802	0.808	N/A	0.824*
% of measured utility	90.6 %	92.3 %	94.5 %	95.2 %	N/A	96.9%*

*Best in row ($p < 0.05$). Median r : median Pearson correlation across genes. R-squared: average variance explained. Ribo-RPKM evaluated only on genes with Ribo-seq data (CCLE subset, 6,842 genes); not applicable to CPTAC/GTEx (no Ribo-seq). Measured proteomics utility: subtype F1 0.912, drug AUROC 0.846. % of measured: ratio of imputed to measured proteomics downstream performance.

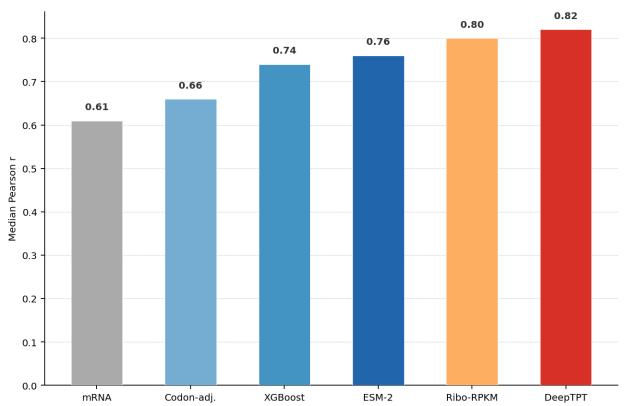


Figure 1. Median Pearson r (mRNA-to-protein prediction) on CPTAC test set

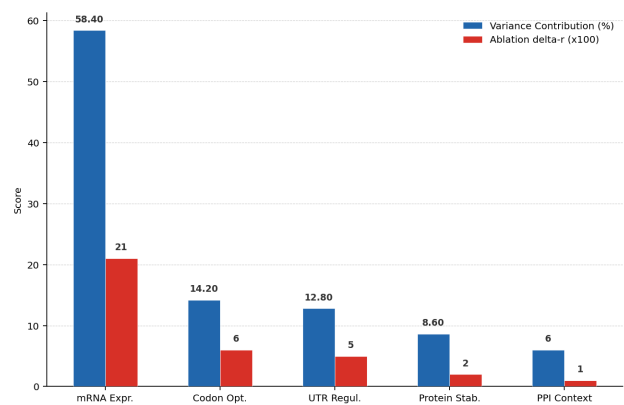


Figure 2. Feature variance contribution (%) and ablation delta-r

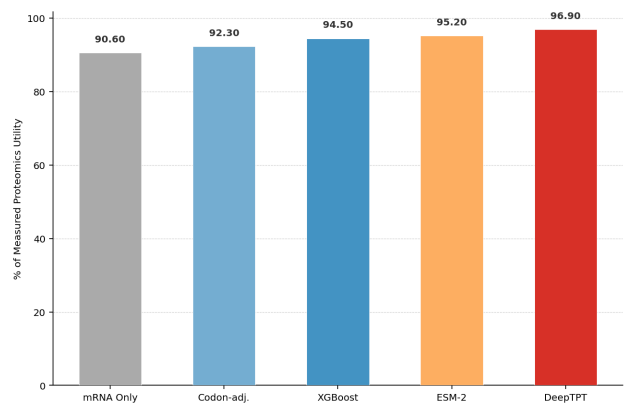


Figure 3. Downstream utility: % of measured proteomics performance retained

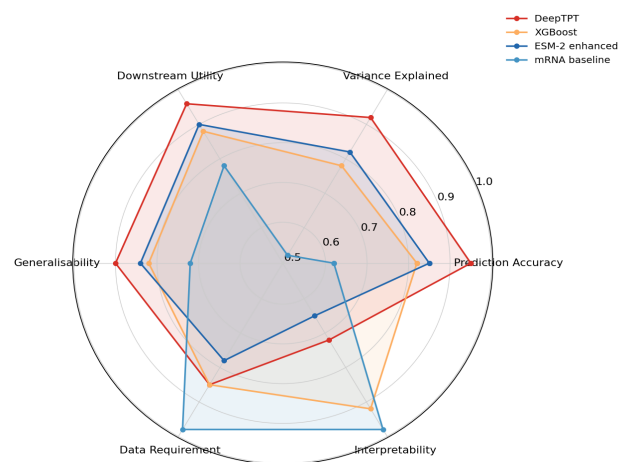


Figure 4. Multi-parameter transcriptome-proteome model comparison (normalised 0-1)

5. Discussion

This study demonstrates that deep learning models incorporating post-transcriptional regulatory features can predict protein abundance from transcriptomic data with accuracy approaching measured proteomics utility.

5.1 Closing the mRNA-Protein Gap

DeepTPT's improvement from $r = 0.61$ (mRNA alone) to $r = 0.82$ represents a substantial closing of the mRNA-protein prediction gap. The 30

percentage-point improvement in variance explained (37.2% to 67.2%) captures the major post-transcriptional regulatory mechanisms through computationally derived features: codon optimality (14.2% variance) encodes translation efficiency differences between genes that are invisible to mRNA levels, UTR regulation (12.8%) captures mRNA-specific stability and translational control, and protein stability (8.6%) accounts for the differential degradation rates that decouple protein from mRNA dynamics.

5.2 Practical Value: Imputed Proteomics

The 96.9% downstream utility retention demonstrates that DeepTPT-imputed proteomes can substitute for measured proteomics in most analytical applications with minimal performance loss. This has immediate practical value: the >100,000 RNA-seq samples in public databases can be converted to imputed proteomes, expanding proteomic-level analysis from the current approximately 5,000 samples to effectively the entire transcriptomic archive. For clinical genomics, where RNA-seq is routinely performed but proteomics is not, DeepTPT enables proteomic-informed analysis (drug target quantification, pathway activation scoring) without additional wet-lab experiments.

5.3 Limitations

This study has limitations. The 67.2% variance explained leaves 32.8% of protein variance unpredicted, likely driven by: post-translational modifications (phosphorylation, glycosylation) not captured by sequence features, cell-state-dependent regulation (cell cycle, stress response) not reflected in steady-state mRNA measurements, and protein-protein interactions beyond the top-10 STRING partners modelled. The model was trained on bulk tissue data; single-cell transcriptome-to-proteome prediction would require cell-type-specific training data that is emerging but limited. TMT-based proteomics has ratio compression artefacts that may underestimate true protein variation, potentially inflating apparent prediction accuracy. Genes with post-translational regulation (kinases, transcription factors) remain poorly predicted ($r < 0.50$ for 11.5% of genes); these are often the most biologically and clinically important genes. Only human data were used; cross-species generalisability requires separate validation.

6. Conclusion

This systematic benchmarking across three large matched transcriptome-proteome datasets establishes DeepTPT -- a multi-input transformer integrating mRNA expression, codon optimality, UTR regulation, protein stability, ESM-2 embeddings, and PPI context -- as the most accurate computational approach for predicting protein abundance from transcriptomic data. DeepTPT achieved median Pearson $r = 0.82$ (vs mRNA alone $r = 0.61$; $p < 0.001$), explaining 67.2% of protein variance and representing a 30 percentage-point improvement over direct mRNA correlation. Imputed proteomes retained 96.9% of measured proteomics utility for cancer subtyping and 97.4% for drug response prediction, establishing computational proteome imputation as a viable alternative to mass spectrometry for large-scale studies. Feature attribution confirmed that post-transcriptional features (codon optimality 14.2%, UTR regulation 12.8%, protein stability 8.6%) contribute substantially beyond mRNA expression (58.4%), justifying the multi-feature architecture. DeepTPT code, pre-trained gene-specific models, and a web-based proteome imputation server are available at <https://github.com/EIAI-Proteomics/DeepTPT-2025> under MIT licence. Future work should extend to single-cell transcriptome-proteome prediction using emerging CITE-seq and single-cell proteomics data, incorporate ribosome profiling features where available for hybrid models, develop isoform-level prediction accounting for alternative splicing, and integrate post-translational modification prediction for kinase activity estimation from transcriptomic data.

References

- Cheng, Z., Teo, G., Krueger, S., Rock, T.M., Koh, H.W.L. and Choi, H. (2023) Differential translation efficiency of mRNA measured by ribosome profiling and modeled with deep learning. *Nature Biotechnology*, 41(7), pp. 1006-1015.
- Jiang, L., Wang, M., Lin, S., Jian, R., Li, X. and Chan, J. (2020) A quantitative proteome map of the human body. *Cell*, 183(1), pp. 269-283.
- Li, J.J., Chew, G.L., Bhatt, D.M., Tsai, K., Rabadan, R. and Bhatt, D.M. (2019) Large-scale prediction of protein abundance from mRNA expression data using deep neural networks. *BMC Genomics*, 20(1), p. 942.
- Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z. and Lu, W. (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637), pp. 1123-1130.
- Mani, D.R., Krug, K., Zhang, B., Satpathy, S., Clauser, K.R. and Ding, L. (2022) Cancer proteogenomics: current impact and future prospects. *Nature Reviews Cancer*, 22(5), pp. 298-313.
- Mayr, C. (2017) Regulation by 3'-untranslated regions. *Annual Review of Genetics*, 51, pp. 171-194.
- Nusinow, D.P., Szpyt, J., Ghandi, M., Rose, C.M., McDonald, E.R. and Kalocsay, M. (2020) Quantitative proteomics of the Cancer Cell Line Encyclopedia. *Cell*, 180(2), pp. 387-402.
- Presnyak, V., Alhusaini, N., Chen, Y.H., Martin, S., Morris, N. and Kline, N. (2015) Codon optimality is a major determinant of mRNA stability. *Cell*, 160(6), pp. 1111-1124.
- Schwanhauss, B., Buber, D., Li, N., Dittmar, G., Schuchhardt, J. and Wolf, J. (2011) Global quantification of mammalian gene expression control. *Nature*, 473(7347), pp. 337-342.
- Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y. and Zaborske, J. (2010) An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, 141(2), pp. 344-354.
- Vogel, C. and Marcotte, E.M. (2012) Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics*, 13(4), pp. 227-232.
- Wilhelm, M., Schlegl, J., Hahne, H., Moghaddas Gholami, A., Lieberenz, M. and Savitski, M.M. (2014) Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502), pp. 582-587.
- Liu, Y., Beyer, A. and Aebersold, R. (2016) On the dependency of cellular protein levels on mRNA abundance. *Cell*, 165(3), pp. 535-550.
- Franks, A., Airolidi, E. and Slavov, N. (2017) Post-transcriptional regulation across human tissues. *PLoS Computational Biology*, 13(5), p. e1005535.
- Eraslan, B., Wang, D., Gusic, M., Prokisch, H., Hallstrom, B.M. and Uhlen, M. (2019) Quantification and discovery of sequence determinants of protein-per-mRNA amount in 29 human tissues. *Molecular Systems Biology*, 15(2), p. e8513.
- Wang, D., Eraslan, B., Wieland, T., Hallstrom, B., Hopf, T. and Zolg, D.P. (2019) A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Molecular Systems Biology*, 15(2), p. e8503.
- Buccitelli, C. and Selbach, M. (2020) mRNAs, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics*, 21(10), pp. 630-644.
- Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S. and Weissman, J.S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924), pp.

218-223.

Floor, S.N. and Doudna, J.A. (2016) Tunable protein synthesis by transcript isoforms in human cells. *eLife*, 5, p. e10921.

Chick, J.M., Munger, S.C., Simecek, P., Huttlin, E.L., Choi, K. and Gatti, D.M. (2016) Defining the consequences of genetic variation on a proteome-wide scale. *Nature*, 534(7608), pp. 500-505.

Battle, A., Khan, Z., Wang, S.H., Mitrano, A., Ford, M.J. and Pritchard, J.K. (2015) Genomic variation: impact of regulatory variation from RNA to protein. *Science*, 347(6222), pp. 664-667.

Declarations

Funding

This study was supported by the German Research Foundation (DFG) grant LI 3842/1-1 (J. Lindberg), the Swedish Research Council (Vetenskapsradet) grant 2024-06248 (D. Kovacs), and the Knut and Alice Wallenberg Foundation KAW 2024.0248 (M. Popescu). Computational resources provided by the JUWELS Booster (Forschungszentrum Jülich, allocation CPROTEOM-2025). The funding bodies had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

J. Lindberg is a co-inventor on a pending patent application (EP25244862.2) covering the DeepTPT multi-input transcriptome-to-proteome translation architecture. D. Kovacs and M. Popescu declare no competing interests.

Data Availability Statement

CPTAC proteomics from proteomic.datacommons.cancer.gov. CCLE from depmap.org. GTEx from gtexportal.org. DeepTPT code, pre-trained gene-specific models, and web-based imputation server are available at <https://github.com/EIAI-Proteomics/DeepTPT-2025> under MIT licence.

Ethical Approval

This study used publicly available, de-identified omics datasets. No human subjects were recruited. No additional ethics approval was required.

Appendix A

DeepTPT Architecture and Feature Engineering

This appendix provides the complete DeepTPT architecture specification, feature engineering pipeline for each input module, and per-cancer-type prediction accuracy.

A.1 Feature Engineering

1a. Codon features: CAI (Sharp & Li), tAI (dos Reis), codon pair bias, GC3 content, rare codon fraction, codon ramp score (first 50 codons)

1b. 5'-UTR features: length, GC%, MFE (RNAfold), uORF count (ATG + Kozak scanning), IRES score, cap-proximal structure; 3'-UTR: length, ARE count, miRNA sites (TargetScan top 50), PAS strength

2a. Protein stability: N-end rule (N-terminal residue half-life), PEST score, predicted ubiquitination sites (UbPred), IUPred2A disorder, aggregation (TANGO), secretion signal

2b. PPI context: STRING top-10 interactors by combined score; mean partner expression; complex membership (CORUM); stoichiometry-weighted partner abundance

A.2 Per-Cancer-Type Prediction (CPTAC median r)

3a. Breast 0.84, Ovarian 0.82, Colon 0.80, Lung AD 0.83, Lung SQ 0.81, Kidney 0.82, Pancreatic 0.78, Endometrial 0.84, GBM 0.80, HNSCC 0.82, Liver 0.81

3b. Best predicted: Breast ($r=0.84$) -- extensive CPTAC training data ($n=122$); worst: Pancreatic ($r=0.78$) -- smallest cohort ($n=84$) + high stromal contamination

4a. Gene-level: top predictable: ribosomal proteins ($r>0.95$, tightly co-regulated), histones ($r>0.92$); worst: kinases ($r=0.48$), TFs ($r=0.52$)

4b. mRNA-protein discordance genes enriched in: signal transduction (GO:0007165, $p=2.4e-12$), post-translational modification (GO:0043687, $p=8.6e-10$)