PharmaronT2P: High-Fidelity Transcriptome-to-Proteome Conversion

Zhe Li, Kaiqiang Hu, Shangxiang Ye, Hui Sun, Wanyu Tao, Yuning Chen, Pengwei Pan, Fang He*

in vitro **Biology, Pharmaron (Beijing) Co., Ltd, Beijing, China** ***Corresponding author: fang.he@pharmaron.com** **bd@pharmaron.com | www.pharmaron.com | poster # 5512**

Abstract

Transcriptome profiling has become a routine, low-cost component of large-scale drug screening. In contrast, direct proteome measurement, while invaluable for assessing drug target engagement and pathway activity at the protein level, remains labor intensive and less accessible, limiting its use in high-throughput screening assays. To fill the gap between transcriptome and proteome, recent studies proposed transcriptome-to-proteome (T2P) prediction models based on public clinical and cell line datasets. However, model performance suffers from limited sample size and lack of comprehensive cell panel coverage, limiting adoption of T2P models in early drug discovery.

Towards T2P models optimized for drug discovery, we generated a foundational dataset of ~800 cancer cell lines with paired transcriptome and proteome profiles using our in-house Orbitrap Astral platform. This internally curated resource is one of the largest and most consistent paired omics datasets to date and forms the basis for accurate T2P prediction.

Here, we developed PharmaronT2P (Pharmaron Transcriptome-to-Proteome), a deep learning framework that maps gene-level RNA expression to protein abundance. The model is pre-trained on public resources (e.g., CPTAC, CCLE) and fine-tuned on our high-quality internal data. PharmaronT2P serves as an accessible *in silico* surrogate for proteomics, converting inexpensive RNA-seq readouts into high-confidence protein signatures. Held-out cell line predictions and GSEA results show the accuracy and potential application of our model.

Together, PharmaronT2P is a valuable resource to guide drug screening assays via *in silico* proteome prediction.

1

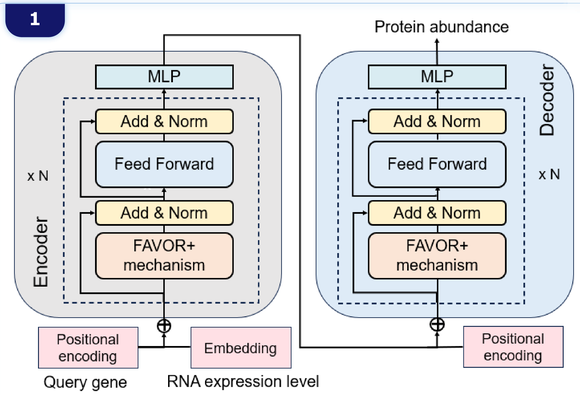

Fig.1 Model Architecture. We employ scTranslator-10K[1] as the pre-trained model, augmented with customized adaptations for transcriptomic and proteomic data. The model takes RNA expression values and their corresponding genes as input. Expression values are projected into fixed-dimension vectors through a learnable embedding layer. In the decoder, positional encodings are derived from the queried protein IDs. The FAVOR+ mechanism enables the model to capture latent long-range dependencies. The model's final output is the predicted relative abundances of the target proteins.

Norm: layer normalization; MLP: multi-layer perceptron

[1] Liu L, Li W, Wang F, et al. A pre-trained large generative model for translating single-cell transcriptomes to proteomes. Nature Biomedical Engineering, 2025.

2

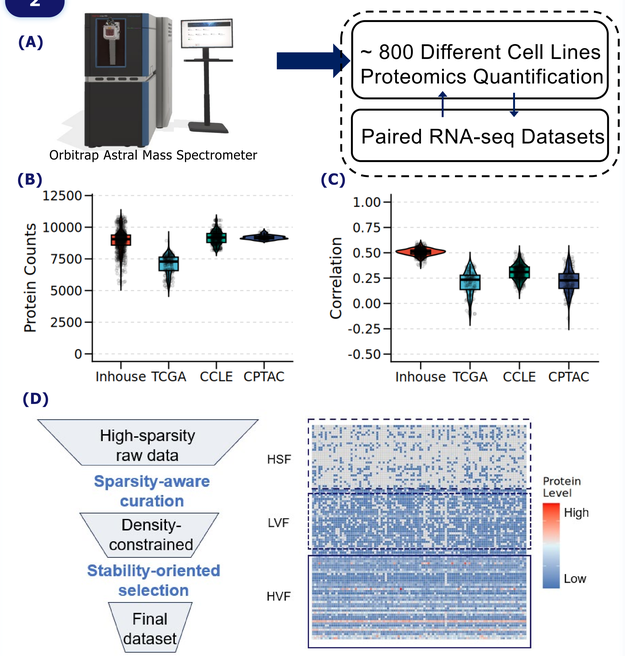

Fig.2 Multi-Source Paired RNA-Protein Dataset Curation for Model Training.

(A) Internal protein quantification platform. We use the Orbitrap Astral Mass Spectrometer for accurate protein quantification with deep coverage and high sensitivity, combined with paired RNA-seq datasets for later fine-tuning.
(B) High-quality internal paired datasets for fine-tuning. Our internal protein quantification results show high coverage compared to public datasets including TCGA and achieved a similarly high level of coverage as CCLE and CPTAC curated set.
(C) Superior RNA-Protein concordance of internal datasets. These internal datasets exhibit strong RNA–protein concordance, with a Spearman correlation of ~0.5, exceeding TCGA, CCLE, and CPTAC references.
(D) Specialized preprocessing for fine-tuning dataset optimization. To ensure effective fine-tuning, we applied additional preprocessing to the fine-tuning datasets, correcting overly sparse regions and adjusting data based on density and variance to optimize downstream performance.

HSF: High-sparsity feature; LVF: Low-variance feature; HVF: High-variance feature.

Highlights

- PharmaronT2P integrates a published pre-trained model with internal fine-tuning, achieving high agreement between predicted and measured protein levels.
- Predicted proteomes better reflect protein-level biology than RNA alone, supported by closer alignment with measured-protein GSEA results.
- We provide an accessible *in silico* alternative to experimental proteomics for drug screening.

3

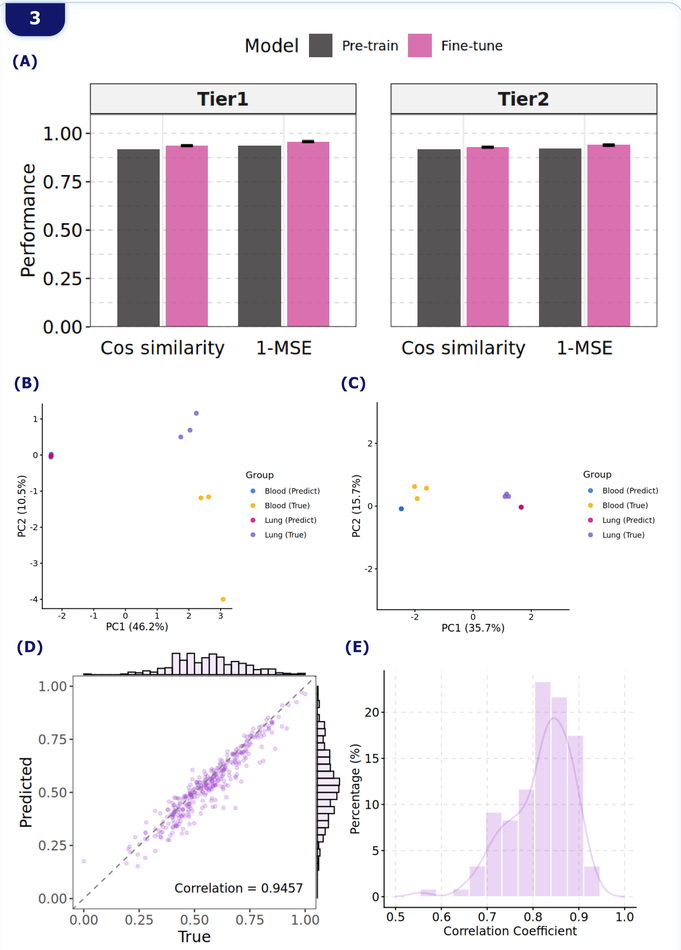

Fig.3 Model Fine-Tuning Performance and Generalization.

(A) Performance on Cancer Gene Census gene sets. Model performance after fine-tuning on internal datasets and COSMIC Cancer Gene Census Tier 1 and Tier 2 gene sets were assessed using five-fold cross-validation.
(B) PCA clustering of measured vs. predicted protein abundance from the pre-trained model. Principal component analysis (PCA) was performed separately on the experimentally measured protein abundance and on the predicted protein abundances from the pre-trained model, showing distinct clustering patterns.
(C) PCA clustering of measured vs. predicted protein abundance of fine-tuned model. Results of fine-tuned model show that predicted values closely recapitulate the tissue-specific clustering pattern of experimental data, verifying that the fine-tuned model retains the feature representation capacity to discriminate between different tissue types.
(D) High concordance between predictions and ground truth. In a representative held-out sample, the fine-tuned model reached a Pearson correlation of 0.9457 between predicted and observed protein levels.
(E) Generalization to held-out cells. We held out 20% of cells from the internal dataset and fine-tuned using the remaining 80%. The fine-tuned model then predicted protein abundances for the held-out cells based on RNA expression. The predictions showed high concordance with the measured protein levels across most held-out cells.

4

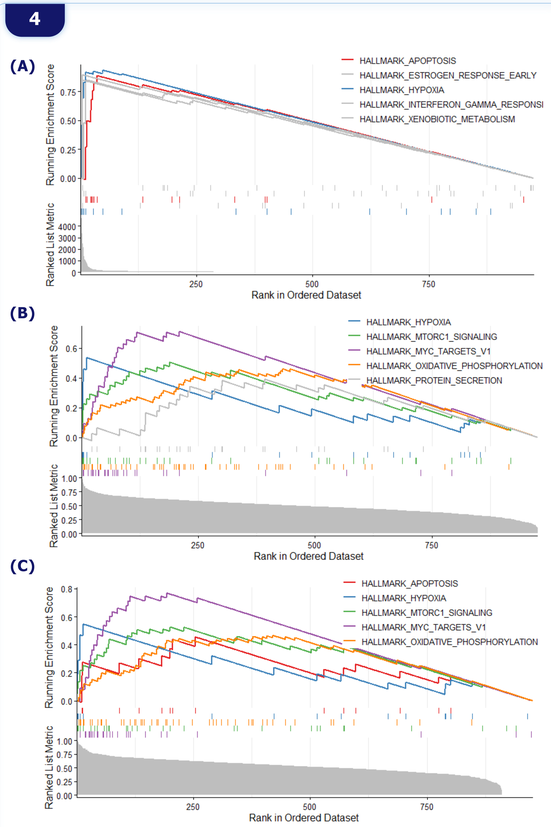

Fig.4 GSEA Comparison Across RNA, Predicted Protein, and Measured Protein Levels.

(A) GSEA based on RNA expression ranking. For the HCT116 cell line, 1,000 randomly selected genes were ranked by RNA expression and analyzed using Hallmark gene sets from MSigDB. The top five enriched pathways were: Apoptosis, Estrogen Response Early, Hypoxia, Interferon Gamma Response, and Xenobiotic Metabolism. These results reflect the transcriptional programs most active at the RNA level.
(B) GSEA based on predicted protein abundance. Using the same gene list, we predicted protein abundance from RNA levels and performed GSEA using rankings based on predicted protein expression. The top five enriched pathways were: Hypoxia, mTORC1 Signaling, MYC Targets, Oxidative Phosphorylation, and Protein Secretion. This pattern suggests that the RNA-to-protein prediction model recovers biological signals that are not evident from RNA alone.
(C) GSEA based on measured protein abundance. Using measured protein levels for ranking, the top five enriched pathways were: Apoptosis, Hypoxia, mTORC1 Signaling, MYC Targets, and Oxidative Phosphorylation. The overlap between predicted-protein–based and measured-protein–based enrichments is substantial, particularly for metabolic and growth-related signatures.
Comparison shows that GSEA results derived from predicted protein levels more closely align with those obtained from measured protein levels than those derived from RNA expression, indicating that the RNA-to-protein prediction better captures underlying biological signals.