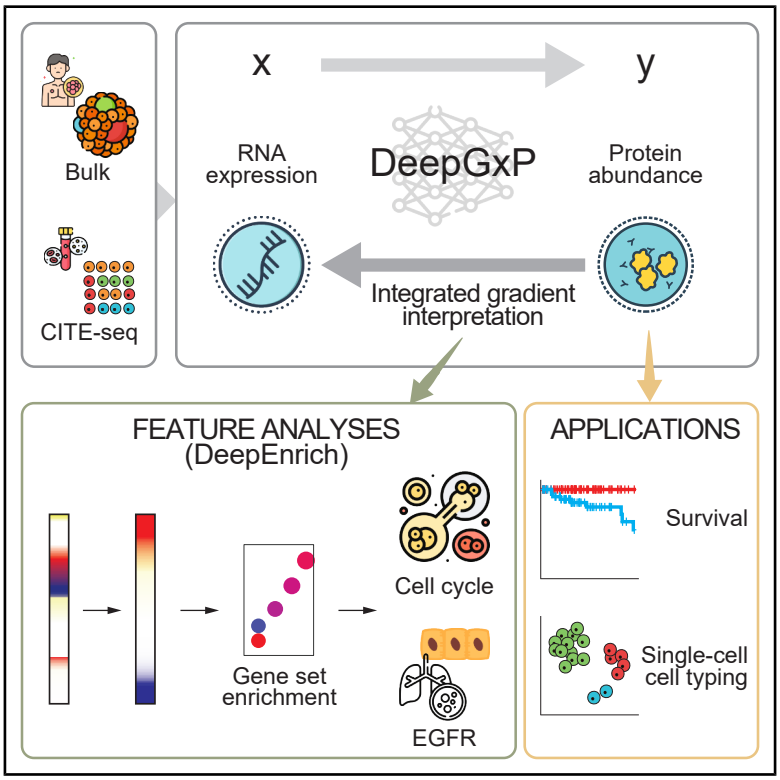
iScience

Article

Predicting and interpreting protein and phosphoprotein abundance from pan-cancer and single-cell transcriptomes

Graphical abstract



Highlights

- DeepGxP provides a framework to translate transcriptomes into proteomic insight
- DeepEnrich links RNA predictors to functional protein pathways and activities
- DeepEnrich reveals cancer type-specific EGFR and HER2 phosphorylation patterns
- Mutation effects are captured even without mutation data in model training

Authors

Hui-Mei Tsai, Tzu-Hung Hsiao, Yu-Chiao Chiu, Yufei Huang, Eric Y. Chuang, Yidong Chen

Correspondence

chuangey@ntu.edu.tw (E.Y.C.), cheny8@uthscsa.edu (Y.C.)

In brief

Protein; Computational bioinformatics; Transcriptomics

Tsai et al., 2026, iScience *29*, 114815
March 20, 2026
https://doi.org/10.1016/j.isci.2026.114815

CellPress

iScience

CellPress
OPEN ACCESS

Article

Predicting and interpreting protein and phosphoprotein abundance from pan-cancer and single-cell transcriptomes

Hui-Mei Tsai,[1,2,3,12] Tzu-Hung Hsiao,[2,4,5,12] Yu-Chiao Chiu,[6,7] Yufei Huang,[6,7] Eric Y. Chuang,[1,8,9,10,*] and Yidong Chen[3,11,13,*]

[1]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei 106319, Taiwan
[2]Department of Medical Research, Taichung Veterans General Hospital, Taichung 407219, Taiwan
[3]Greehey Children's Cancer Research Institute, University of Texas Health San Antonio, San Antonio, TX 78229, USA
[4]Department of Public Health, Fu Jen Catholic University, New Taipei 24205, Taiwan
[5]Institute of Genomics and Bioinformatics, National Chung Hsing University, Taichung 402202, Taiwan
[6]Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA
[7]UPMC Hillman Cancer Center, University of Pittsburgh, Pittsburgh, PA 15232, USA
[8]Biomedical Technology and Device Research Laboratories, Industrial Technology Research Institute, Hsinchu 310401, Taiwan
[9]Research and Development Center for Medical Devices, National Taiwan University, Taipei 106319, Taiwan
[10]Department of Electrical Engineering, College of Electrical Engineering and Computer Science, National Taiwan University, Taipei 106319, Taiwan
[11]Department of Population Health Sciences, University of Texas Health San Antonio, San Antonio, TX 78229, USA
[12]These authors contributed equally
[13]Lead contact
*Correspondence: chuangey@ntu.edu.tw (E.Y.C.), cheny8@uthscsa.edu (Y.C.)
https://doi.org/10.1016/j.isci.2026.114815

SUMMARY

Proteins that impact phenotype and disease are often approximated by RNA expression, which poorly infers protein abundance. We developed DeepGxP, a deep-learning model trained on The Cancer Genome Atlas pan-cancer data, to predict protein abundance from transcriptome profiles. DeepGxP outperformed conventional models, achieving median Pearson's correlation of 0.68 (n = 187) and predictive performance of 0.74 and 0.64 for proteins with high (≥0.31) and low (<0.31) self-gene/protein correlation, respectively. We also developed DeepEnrich, an integrated gradient-based interpretation framework that identifies predictor genes and enriched functions. For example, predictors of cyclin B1 and E2 are enriched in mitotic chromatid segregation and G2/M transition, respectively. In lung adenocarcinoma, we uncovered distinct EGFR/HER2 phosphorylation patterns in alveolar cells. In breast cancer, p53 protein, but not *TP53* mRNA, correlated with survival. DeepGxP also accurately predicted the abundance of single-cell surface proteins, confirming cell identification. Our findings underscore DeepGxP's potential in decoding gene-to-protein relationships for cancer biomarker discovery.

INTRODUCTION

Gene expression profiling has been widely used in cancer research for classification, subtyping, biomarker discovery, therapeutic target identification and predicting drug response, cancer progression, and survival.[1–8] Large-scale transcriptomic data motivated the development of enrichment analyses,[9] which decipher the biological significance of selected gene sets and link gene expression to functional pathways and disease mechanisms.

Although gene expression should directly reflect protein abundance, per-gene RNA-protein correlations remain modest, typically ranging from ∼0.3 to 0.4[10–14] in early studies to ∼0.5[15–17] in recent tumor-specific analyses. This discordance challenges the use of mRNA as a direct proxy for protein abundance and the assumption of a direct, one-to-one relationship between transcriptomic and proteomic data. Two distinct correlation measures are commonly reported in transcriptome-proteome studies: within (per)-gene correlation, which captures a gene's own mRNA-protein relationship across samples, and cross-gene correlation, which assesses how different mRNA and protein expression of genes vary within a sample.[12,18–25] In this study, we focus on within-gene correlations, as they more directly evaluate the predictive power of RNA for protein abundance across samples.

While technical noise contributes to this discrepancy, this mRNA-protein discordance is mainly due to biological complexity, including spatial and temporal variation, biosynthesis and degradation rates, and multi-gene regulation of a protein.[12,21] To address this discordance, several studies have

attempted to predict protein abundance from transcriptomics data. Early efforts employed conventional machine learning methods, such as support vector regression,[26] random forest,[27] and random forest-based ensemble models,[28,29] while more recent work has introduced deep learning architectures such as residual neural network.[30] These models were trained across diverse datasets, ranging from targeted canine data to human data from Clinical Proteome Tumor Analysis Consortium (CPTAC), which profiles up to ~10,000 proteins across multiple tumor types; however, even the largest multi-tumor prediction models have been limited to eight cancers. Beyond predictive accuracy, interpretability has also been explored, for example, Srivastava et al.[27] and Cranney and Meyer[30] using SHapley Additive exPlanation (SHAP) analysis to demonstrate the influence of interacting partner transcripts and identify predictive features. However, most prior models remain constrained by dataset scope or methodological focus. Moreover, conventional machine learning methods may be insufficient to capture nonlinear relationships.[31] These gaps highlight the need for a pan-cancer, interpretable deep learning framework to systematically dissect the determinants of protein abundance from transcriptomic data.

Protein abundance in clinical practice is primarily measured by antibody-based assays such as immunohistochemistry, which are widely applied for disease diagnosis, inferring survival, and drug responses,[32,33] whereas mass spectrometry, Olink, and SomaScan are mainly used in research. Proteins serve as direct functional molecules and are the most common drug targets and biomarkers in clinical settings. The proteome provides distinct molecular signatures that reflect tumor subtypes and patient outcomes, distinct from those identified by the transcriptome.[16] Recent large-scale studies further demonstrate that plasma proteomic signatures can robustly predict the incidence of a wide spectrum of common and rare diseases, including hematological, cardiovascular, respiratory, and autoimmune conditions, underscoring the clinical utility of protein-based predictors beyond transcriptomic associations.[34]

Disease-associated expression changes inferred from predicted protein abundances are not evident at the mRNA level, as seen in Alzheimer's disease,[35] highlighting the value of proteomics in understanding disease mechanisms beyond transcriptional control. Using a high-throughput antibody-based method, similar to immunohistochemistry, reverse-phase protein assays (RPPAs) profiled ~200 total and phosphorylated proteins in The Cancer Genome Atlas (TCGA) samples, at the time of analysis.[11,36,37] RPPA has enabled tumor classification, cancer signaling analysis, and biomarker discovery,[38–40] proving its effectiveness in deciphering tumor biology.

Deep learning is a branch of artificial intelligence that uses intricate neural networks capable of learning complex features from big data. The accumulation of genomic profiles of cell lines and tumor samples has enabled the use of deep learning models for predicting cancer type, drug response, and survival, as well as binding motifs and variant effects from DNA sequences.[41–44] The convolutional neural network (CNN) is one of the most well-developed deep learning models for omics data; implementation focuses mostly on classification problems, such as cancer-type[45,46] and survival[47] predictions. CNN models are flexible with regard to input shape. For example, Mostavi et al.[45] and Lyu et al.[46] proposed 1- and 2-dimensional mapping of gene expression profiles using TCGA data; all achieved >95% prediction accuracy for 33 cancer types. CNNs have also modeled genotype-phenotype associations in synthetic biology, particularly in predicting protein abundance from DNA sequences for selecting high-yield strains in soybeans[48] and *E. coli*.[49,50] CNN and generative adversarial networks have also shown promise to help understand how genomic regulatory structures contribute to RNA expression.[51,52]

To address the issue of predicting protein abundance from the transcriptome and expand the utility of gene expression profiling, we developed **Beep** learning for **G**ene e**x**pression mapping to **P**rotein abundance (DeepGxP), a 1-dimensional CNN model that learns the translational process from high-dimensional transcriptomic data. We trained the model on data from TCGA, validated it using data from the independent International Cancer Genome Consortium (ICGC), and extended it to recent single-cell transcriptomics datasets. Given the challenge of predicting protein abundance, we chose antibody-based RPPA data instead of mass spectrometry-based data from CPTAC, leveraging its larger sample size and lower feature-to-sample ratio. Unlike previous models that prioritize protein prediction accuracy, DeepGxP focuses on providing insights into the regulatory mechanisms that shape protein abundance. The model represents a steady-state relationship between the transcriptome and the proteome, and temporal effects can be minimized.[25] Our primary goal in using DeepGxP is to enable biological interpretations rather than pure predictive performance.

To enhance interpretability, we introduce DeepEnrich, a framework that identifies protein predictors, and associated molecular pathways and functions that influence protein abundance. DeepEnrich enables the discovery of transcriptional and post-transcriptional regulatory mechanisms. Together, DeepGxP and DeepEnrich bridge transcriptomics and proteomics, moving beyond correlation-based approaches and providing a systems-level framework to study protein regulation in cancer biology.

RESULTS

DeepGxP model performance

DeepGxP utilizes deep learning models to determine protein abundance through whole-transcriptome expression profiles. A CNN model was constructed and trained with 13,995 genes from 6,813 solid tumors across 27 cancer types in the TCGA cohort to predict 187 proteins. These proteins were selected based on consistency between RPPA and mass spectrometry measurements within the Cancer Cell Line Encyclopedia (CCLE) dataset, ensuring reliable quantification across platforms (see STAR Methods). The detailed model architecture is depicted in Figure 1A. DeepGxP achieved an overall mean square error in protein abundance prediction of 0.086 ± 0.005, 0.190 ± 0.001, and 0.187 ± 0.002 in training, validation, and testing data, respectively. In the testing set, the median Pearson's correlation of prediction was 0.68 across 187 proteins, significantly improved (one-sided Mann-Whitney $p = 3.78 \times 10^{-36}$) compared to the median correlations of 0.31 between self-gene expression

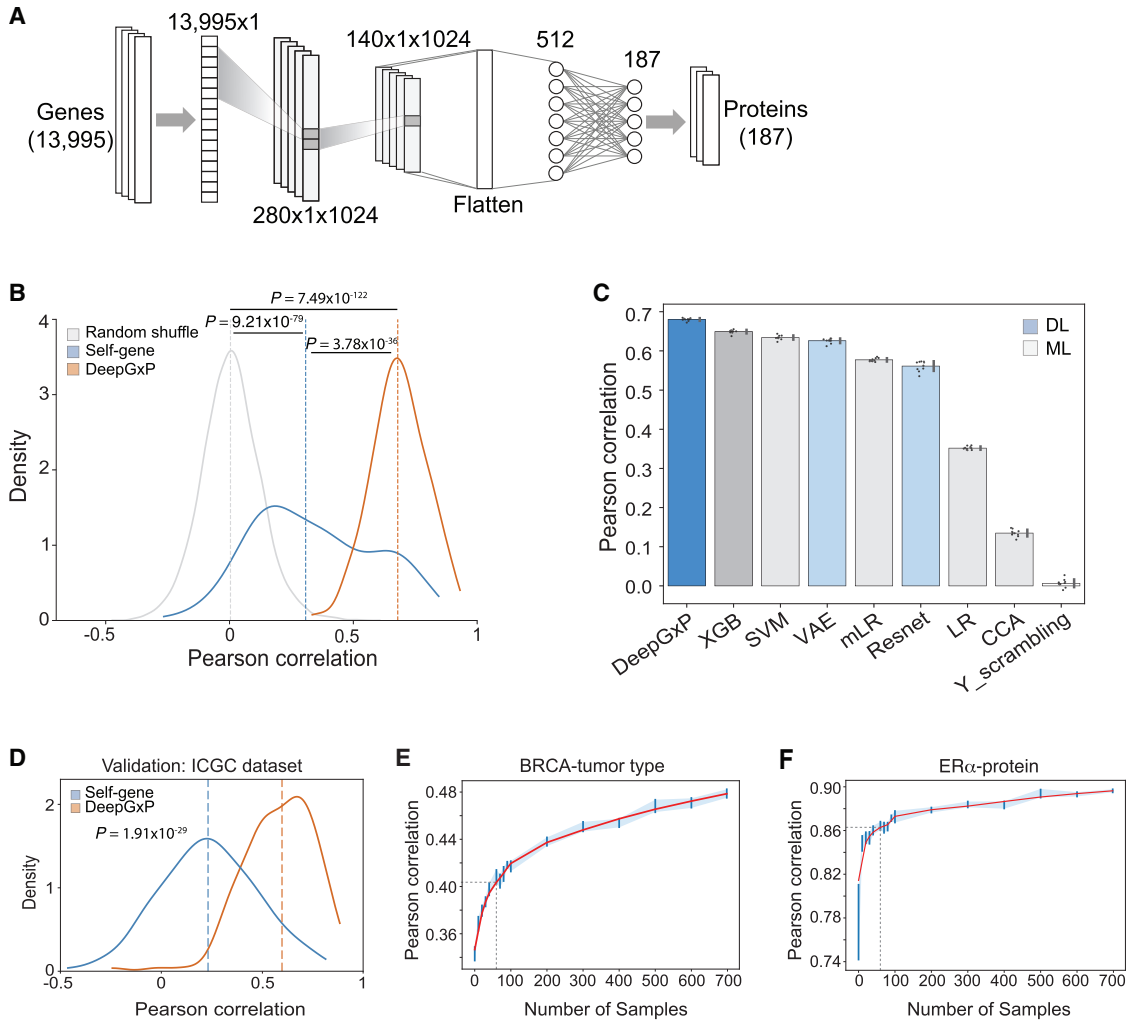


Figure 1. Performance and validation of DeepGxP

(A) DeepGxP model structure, which includes one convolution layer and one max pooling layer, followed by one dense fully connected layer and output layers with 187 nodes for abundance prediction of 187 proteins.

(B) Overall performance of DeepGxP. Per-protein Pearson's correlations shown between random mRNA-protein pairs (gray), corresponding self-gene/protein pairs (blue), and DeepGxP-predicted vs. real protein pairs (orange). The mean correlations are shown as dashed lines, which are 0.00 (9,350 pairs), 0.31 (187 pairs), and 0.68 (187 pairs), respectively. Statistical significance was tested with the one-tailed Mann-Whitney test.

(C) Performance comparisons with other deep learning (VAE and ResNet) and machine learning methods. Pearson's correlation coefficients were derived from the average per-protein Pearson's correlation from 10-fold cross-validation. Each dot represents the result from the testing set after each fold of training. Data shown as mean ± standard deviations.

(D) Applying DeepGxP on the independent ICGC dataset. Per-protein Pearson's correlation between protein abundance and DeepGxP-predicted (brown) and self-gene expression (blue). The median correlations are shown as dashed lines (0.60 and 0.23, respectively). Statistical significance was determined using a one-tailed Wilcoxon signed-rank test.

(E and F) Robustness of DeepGxP in learning unseen data. DeepGxP was initially trained without any BRCA samples. Performance on (E) average overall proteins and (F) a single estrogen receptor-α protein was assessed on the breast cancer-only test set after a consecutive number of breast cancer samples were trained. The dashed line depicts the turning point when the training sample size reached 60. VAE, variational autoencoder; XGB, XGBoost; SVM, support vector machine; LR, univariate linear regression; mLR, multivariate linear regression; CCA, canonical correlation analysis.

and protein abundance (ρ_{self}) (Figure 1B). DeepGxP predictions were compared against self-gene mRNA expression since mRNA is often used as a proxy for protein abundance. To ensure model robustness and prevent overfitting, we performed 10-fold cross-validation against a collection of deep learning and machine learning models. DeepGxP has the highest mean Pearson's correlation (0.680 ± 0.004) (Figure 1C and Table S3) and significantly outperformed other models, as assessed by paired one-tailed *t* test (extreme gradient boosting (XGBoost) [XGB]: $p = 1.31 \times 10^{-10}$, support vector machine [SVM]: $p = 8.14 \times 10^{-9}$, variational autoencoder [VAE]: $p = 1.67 \times 10^{-10}$, multivariate linear regression [mLR]: $p = 8.69 \times 10^{-15}$, residual network

[ResNet]: $p = 2.41 \times 10^{-11}$, LR: $p = 4.82 \times 10^{-18}$, canonical correlation analysis [CCA]: $p = 3.83 \times 10^{-17}$, and *y*-scrambling: $p = 2.64 \times 10^{-18}$). Compared to the XGB method (the second-best model), DeepGxP achieved a 13.1% improvement in prediction accuracy. To further validate DeepGxP and assess generalizability, we tested it on an independent dataset from the ICGC ($n = 3,053$, predicting 169 out of 187 proteins). The predicted protein abundances correlated well with assayed protein abundance (median $\rho = 0.60$ across 169 matched proteins) (Figure 1D), confirming that DeepGxP was not overfit to the TCGA training data. Moreover, DeeGxP also had good performance when RPPA and mass spectrometry measurements were concordant (Figure S1).

We also assessed the ability of the DeepGxP model to handle heterogeneous tumors with each cancer type, in terms of the minimum number of samples required to achieve a robust estimation of protein abundance. To achieve this goal, we first retrained DeepGxP without breast cancer (BRCA) samples and incrementally added them back. The model's performance was evaluated using a 10% testing sample set aside before the training procedure. As shown in Figure 1E and Table S4, the initial estimates of protein abundance (DeepGxP model trained without any BRCA samples) only achieved a Pearson's correlation of 0.34 compared to actual measurements. The coefficient increased to 0.48 when all 698 BRCA samples were used. We determined that when 60 samples were used, the coefficient reached the turning point, or $\rho = 0.42$ for all proteins (Figure 1E) and 0.86 for estrogen receptor-α (Figure 1F).

Accuracy of DeepGxP in predicting abundance of phosphoproteins

We investigated the ability of DeepGxP to predict self-gene expression (protein with corresponding RNA) and then phosphoprotein abundance. First, we evaluated the correlation difference between DeepGxP and self-gene prediction. For proteins with low self-gene/protein correlation (low self-gene correlation [sGP], $\rho_{self} < 0.31$), DeepGxP significantly improved prediction performance, increasing the median correlation coefficient from 0.14 (self-gene prediction) to 0.64 (DeepGxP prediction, $\rho_{DeepGxP}$) (one-tailed Wilcoxon signed rank $p = 1.90 \times 10^{-17}$), with a median $\Delta\rho$ ($\rho_{DeepGxP} - \rho_{self}$) of 0.46 (Figure 2A). For proteins with high self-gene/protein correlation (high-sGP, $\rho_{self} \geq 0.31$), DeepGxP also improved prediction performance significantly, from 0.54 (self-gene) to 0.74 (DeepGxP) (one-tailed Wilcoxon signed rank $p = 2.79 \times 10^{-17}$; median $\Delta\rho$ of 0.18). Improvements in prediction performance differed significantly between the low-sGP and high-sGP groups (one-tailed Mann-Whitney test, $p = 2.38 \times 10^{-28}$). For example, estrogen receptor-α, a high-sGP protein ($\rho_{self} = 0.86$), showed an additional 8.1% improvement from DeepGxP prediction ($\rho_{DeepGxP} = 0.93$, Figure 2B). On the other hand, HSP70, a low-sGP protein ($\rho_{self} = 0.07$), showed a marked improvement of 7.29-fold ($\rho_{DeepGxP} = 0.58$, Figure 2C).

To assess whether the prediction has a tumor-specific pattern, we evaluated DeepGxP performance in BRCA and lung adenocarcinoma (LUAD). Most proteins (86.6%, 162/187) showed consistent predictive performance across both cohorts, while 10.2% (19/187) and 3.2% (6/187) were BRCA- and LUAD-specific, respectively (Figure S2). For example, GATA3 was well predicted in BRCA ($\rho = 0.86$) but not LUAD ($\rho = 0.02$), reflecting its high abundance and variability in BRCA (mean, $\rho = 1.57$, standard deviation = 1.15) but absence in LUAD (mean $\rho = -0.19$, standard deviation = 0.22). Conversely, phosphorylated SHP2 at Y542 (pSHP2-Y542) was well predicted in LUAD ($\rho = 0.64$) but not BRCA ($\rho = 0.21$). Cyclin B1, by contrast, showed consistent predictability in both tumor types ($\rho = 0.84$ in both), reflecting conserved cell cycle control.

Phosphoproteins are less correlated with expression of self-genes compared to total proteins, with only 11/56 (~20%) in the high-sGP group. However, DeepGxP significantly improved prediction performance for phosphoproteins compared to total proteins (median $\Delta\rho_{Total} = 0.26$, median $\Delta\rho_{Phospho} = 0.46$, two-tailed Mann-Whitney $p = 7.25 \times 10^{-9}$, Figure 2D). Phosphoprotein prediction in the low-sGP group (median $\Delta\rho = 0.49$) also significantly improved compared to the high-sGP group (median $\Delta\rho = 0.18$) (one-tailed Mann-Whitney $p = 9.49 \times 10^{-7}$). The results were consistent in BRCA and LUAD cohorts (Figure S2B). These results indicate that DeepGxP improved phosphoprotein predictions, likely due to the model's ability to capture the effects of complex phosphorylation regulation.

To examine whether protein abundance prediction was related to their functions, we further partitioned proteins according to their function classification.[53] Total proteins involved in RNA metabolism and translation demonstrated the greatest increase in correlation (Figure 2E), despite having the lowest self-gene correlations in our dataset (median $\rho = 0.20$ for RNA metabolism and 0.23 for translation, Figure S3). A previous study[14] showed a low copy number-protein correlation but a high copy number-RNA correlation in these groups, likely reflecting predominant regulations driven by post-transcriptional and translational mechanisms occurring after transcription. Their predictive improvement suggests indirect regulatory effects captured in the transcriptome, such as contributions from upstream regulators, co-expressed genes, gene-gene interactions, and RNA-binding proteins. In contrast, total proteins of transmembrane receptors and calcium-binding proteins exhibited the least increases, possibly due to their localization on the cell membrane and their roles at the initiation points of signaling pathways. Despite the smaller number of phosphoproteins and their annotated protein classes, phosphoproteins showed greater gains in prediction performance than total proteins in every functional group where they are present, reflecting the inability of self-gene mRNA alone to capture phosphorylation events. These results illustrate the ability of DeepGxP to estimate phosphorylation and expression of proteins that undergo complex post-transcriptional regulation and exhibit weak direct correlation with their corresponding gene.

DeepGxP model interpretation via DeepEnrich with integrated gradient score and enrichment analyses

We next developed DeepEnrich, a framework for evaluating gene attribution in protein abundance prediction through integrated gradient and subsequent enrichment analyses on the identified positive/negative leading EdGe (pLEG/nLEG) predictors (see STAR Methods and Figure S4). DeepEnrich incorporates two methodologic steps to ensure robust identification of

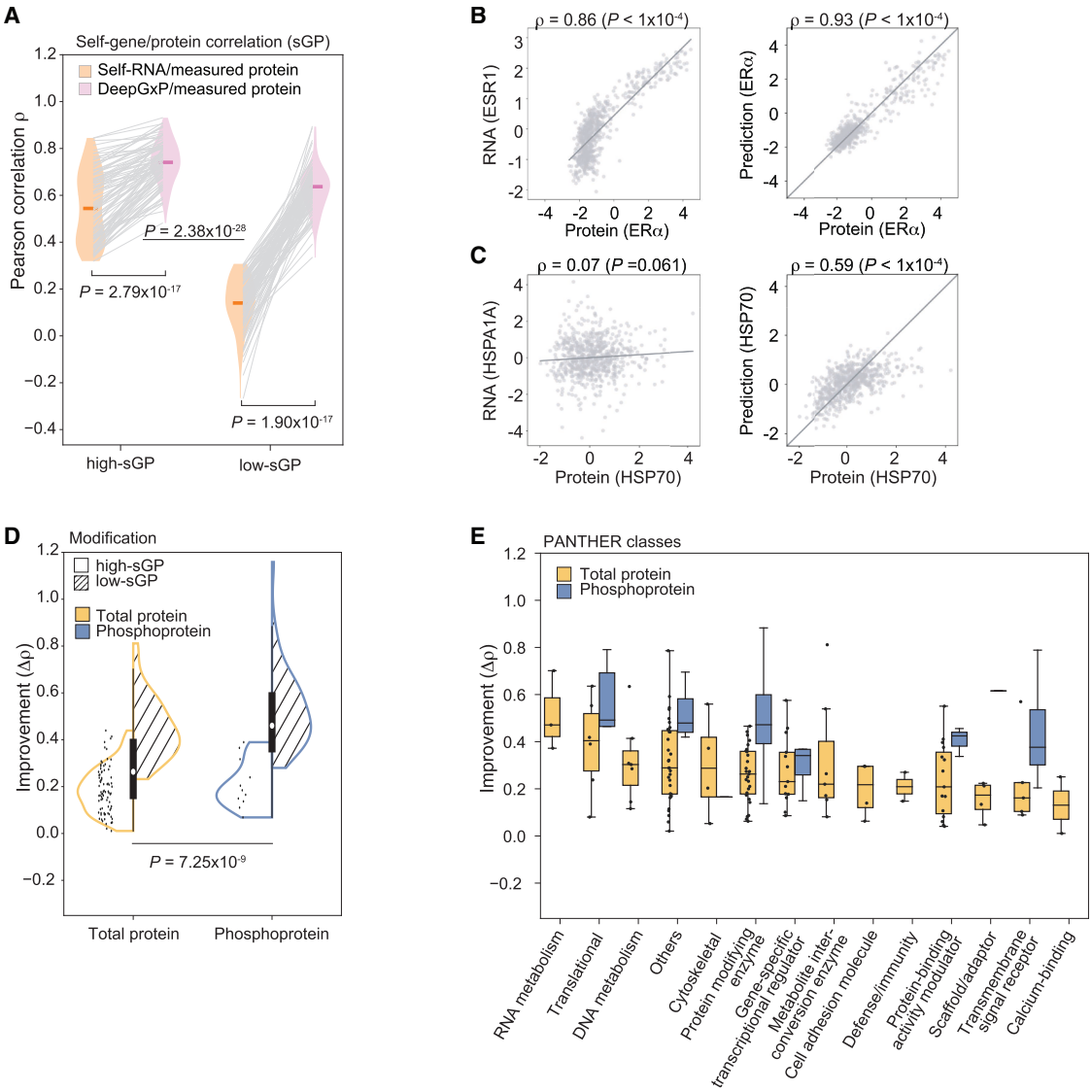


Figure 2. Assessment of protein features underlying prediction improvement

(A) Improvements in prediction based on high ($\rho_{self} \geq 0.31$) and low ($\rho_{self} < 0.31$) self-gene correlation (high/low-sGP). Statistical significance was calculated with one-tailed Wilcoxon signed-rank tests between self-RNA and DeepGxP, while one-tailed Mann-Whitney tests were used to compare high- and low-sGP.

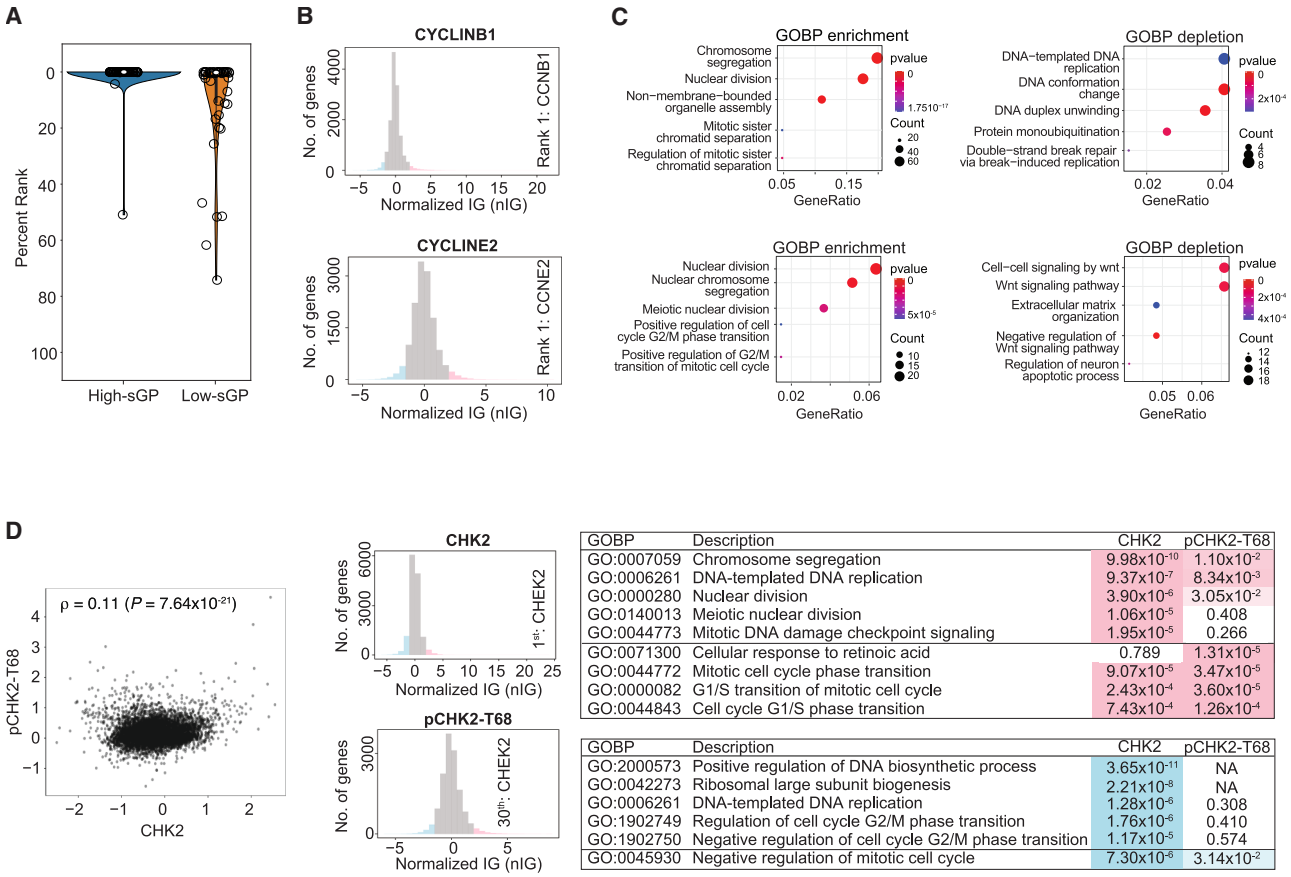
(B and C) Performance in assessing high- and low-self-gene correlated proteins, estrogen receptor-α, and HSP70, respectively. Ground truth and self-gene expression (left) and ground truth and DeepGxP-predicted protein abundance (right).

(D) Comparison of improvements in prediction accuracy based on total or phosphoproteins. Statistical significance was calculated with two-tailed Mann-Whitney tests.

(E) Comparison of improvements in prediction among PANTHER protein classes.

predictors: (1) averaging integrated gradient values from the top 10% protein-expressed samples to minimize noise from lowly expressed samples and (2) normalizing integrated gradient values (Equation 3 in STAR Methods) with stringent thresholds (±1.96), which corresponds to $p < 0.05$ for pLEG and nLEG predictor definition (assuming a normal distribution).

To evaluate the effectiveness of DeepEnrich in identifying biologically protein-specific genes, we first focused on total protein prediction. We will discuss more complex phosphoproteins in the functional enrichment analyses of phosphorylated proteins section. In the high-sGP group, as expected, the normalized integrated gradient of self-genes for the corresponding protein consistently ranked high. Nearly 94% (77/82) of self-genes ranked within the top 10% of input genes, and 83% (68/82) ranked first (Figure 3A and Table S5). For example, cyclin B1, which strongly correlates with its self-gene *CCNB1* (ρ_{self} = 0.84, $\rho_{DeepGxP}$ = 0.91), had the highest normalized integrated gradient, indicating its dominant contribution to cyclin B1 abundance over regulations from other genes (Figure 3B, top). In the low-sGP group, 60% (29/48) of self-genes ranked in the top 10%



GOBP	Description	CHK2	pCHK2-T68
GO:0007059	Chromosome segregation	9.98×10^{-10}	1.10×10^{-2}
GO:0006261	DNA-templated DNA replication	9.37×10^{-7}	8.34×10^{-3}
GO:0000280	Nuclear division	3.90×10^{-6}	3.05×10^{-2}
GO:0140013	Meiotic nuclear division	1.06×10^{-5}	0.408
GO:0044773	Mitotic DNA damage checkpoint signaling	1.95×10^{-5}	0.266
GO:0071300	Cellular response to retinoic acid	0.789	1.31×10^{-5}
GO:0044772	Mitotic cell cycle phase transition	9.07×10^{-5}	3.47×10^{-5}
GO:0000082	G1/S transition of mitotic cell cycle	2.43×10^{-4}	3.60×10^{-5}
GO:0044843	Cell cycle G1/S phase transition	7.43×10^{-4}	1.26×10^{-4}

GOBP	Description	CHK2	pCHK2-T68
GO:2000573	Positive regulation of DNA biosynthetic process	3.65×10^{-11}	NA
GO:0042273	Ribosomal large subunit biogenesis	2.21×10^{-8}	NA
GO:0006261	DNA-templated DNA replication	1.28×10^{-6}	0.308
GO:1902749	Regulation of cell cycle G2/M phase transition	1.76×10^{-6}	0.410
GO:1902750	Negative regulation of cell cycle G2/M phase transition	1.17×10^{-5}	0.574
GO:0045930	Negative regulation of mitotic cell cycle	7.30×10^{-6}	3.14×10^{-2}

Figure 3. Predictor genes for total and phosphorylated cell cycle-related genes captured by DeepGxP are functionally enriched in cell cycle terms

(A) The rank of self-genes as predictors for their corresponding total proteins in percentage (the full list is in Table S5).

(B) Distribution of normalized IGs for (top) cyclin B1 and (bottom) cyclin E2. Both *CCNB1* and *CCNE2* are ranked as the first predictor in their corresponding protein prediction.

(C) Top 5 GOBP terms enriched or depleted in pLEG (left) and nLEG (right), respectively. The full lists of enriched/depleted terms of cyclin B1 and E2 are in Tables S6 and S7, respectively.

(D) CHK2 and pCHK2-T68 protein abundance (left). Distribution of normalized IGs (middle) for CHK2 (top) and pCHK2 (bottom), where its self-gene *CHEK2* ranked first as a predictor for total protein prediction but only 30th for pCHK2-T68 abundance prediction. Comparison of the top 5 (top) enriched and (bottom) depleted GOBP terms in the pLEG and nLEG of CHK2 and pCHK2 (right). The last 2 columns show the enrichment *p* values (Fisher's exact test). Pink and blue shading indicate enriched and depleted terms, respectively (at $p < 0.05$).

of input genes, and 31% (15/48) remained ranked as first (Table S5). Despite the weak self-gene correlation of cyclin E2 with *CCNE2* (ρ_{self} = 0.21, $\rho_{DeepGxP}$ = 0.60), *CCNE2* still ranked first (Figure 3B, bottom), consistent with previous studies, suggesting a noteworthy contribution to cyclin E2 abundance. Interestingly, aside from self-genes, the pLEGs of cyclin B1 and E2 also included other cyclin family members, *CCNE1*, *CCNB2*, *CCNA1*, and *CCND1* for cyclin B1; and *CCNB1*, *CCNE1*, *CCNF*, *CCNA2*, *CCND1*, *CCNB2*, *CCNA1*, and *CCNDBP1* for cyclin E2. Several genes (*CCNE1*, *CCND1*, *CCNB2*, and *CCNA1*) emerged as predictors for both proteins, indicating shared regulatory networks. This convergence validates the biological coherence of our predictive framework, demonstrating that the models successfully captured genuine cell cycle regulatory relationships.

To explore the pathways and functional associations between pLEG/nLEG predictors and protein abundance, we performed enrichment analyses of pLEG and nLEG predictors. Cell cycle-related processes were enriched in pLEG predictors of cyclin B1 and cyclin E2 (Figure 3C, left). Enrichment in cyclin B1 was in GO:0000070: mitotic sister chromatid segregation ($p = 4.48 \times 10^{-35}$), consistent with its peak abundance during G2/M phase.[54] Enrichment in cyclin E2 was in GO:0010971: positive regulation of G2/M transition of mitotic cell cycle ($p = 3.11 \times 10^{-5}$), likely reflecting aberrant extension of cyclin E2 expression into G2 phase in tumors instead of being confined to S phase as in normal cells.[55] nLEG predictors for cyclin B1 were depleted in DNA replication terms, consistent with previous findings that cyclin B1 starts to gradually accumulate during S phase when DNA replication occurs, and its depletion triggers

DNA re-replication. Conversely, nLEG predictors of cyclin E2 were depleted in Wnt signaling pathways (Figure 3C, right), aligning with previous evidence that cyclin E-CDK2 promotes β-catenin phosphorylation and degradation, thereby directly linking cyclin E to Wnt signaling.[56,57] Reactome enrichment and depletion showed similar results (Tables S6 and S7).

For phosphorylated proteins, regulatory involvement of kinases and phosphatases is anticipated. Consistent with this expectation, analysis of pLEG for phosphorylation at Y1068 (pEGFR-Y1068) revealed significant enrichment in GO:0043405, regulation of MAP kinase activity ($p = 1.30 \times 10^{-5}$). The pLEG overlapped with the term includes kinases (*EGFR*, *ERBB2*, *PRKCD*, and *KSR1*) and phosphatases (*DUSP1* and *DUSP7*).

These results underscore the interpretability of DeepGxP at the pathway or functional levels, revealing pathways through enrichment analyses of genes in pLEG (or nLEG) to predict specific proteins with the DeepEnrich method. The full lists of enriched pathways for cyclin B1 in Table S6 and cyclin E2 in Table S7. All other proteins are provided in Tables S8 and S9 for pLEG and nLEG, respectively.

Functional enrichment analyses of phosphorylated proteins

Changes in phosphorylation levels lead to altered protein function and signaling, but do not necessarily correlate with total protein abundance. For instance, the Pearson's correlation between CHK2 total protein and its phosphorylated form at T68 (pCHK2-T68) was only 0.11 ($p = 7.64 \times 10^{-21}$, Figure 3D). Nonetheless, *CHEK2* ranked as the top pLEG predictor for total CHK2 protein and 30th for pCHK2-T68 (Figure 3D, middle).

To elucidate phosphorylation-specific pathways, we used pCHK2-T68 specific predictors for enrichment analyses. The result revealed distinct biological processes: CHK2 was enriched in cell division ($P \sim 1.06 \times 10^{-5} - 9.98 \times 10^{-10}$), while pCHK2-T68 was associated with the G1/S transition ($P \sim 3.6 \times 10^{-5} - 1.26 \times 10^{-4}$, Figure 3D, top right). Conversely, functional depletion analyses indicated that CHK2 was depleted in G2/M transition ($P \sim 1.17 \times 10^{-5} - 1.76 \times 10^{-6}$), and pCHK2-T68 in the overall cell cycle (Figure 3D, bottom right). These findings highlight the unique roles and regulatory mechanisms of phosphorylated versus total proteins found in DeepGxP through DeepEnrich. A full list of enriched pathways for 36 phosphoproteins with corresponding total proteins is provided in Tables S10 and S11 for pLEG and nLEG, respectively.

Phosphorylation-site-specific regulation of EGFR and HER2 signaling in LUAD.

We investigated the total protein abundance of EGFR (tEGFR) and its phosphorylation (p) sites at Y1068 and Y1173 (pEGFR-Y1068 and pEGFR-Y1173), which induce different signaling pathways and functional activities in LUAD.[58] Correlations of tEGFR with pEGFR-Y1068 and pEGFR-Y1173 were moderate (Figure S5A). The phospho-specific pLEGs in pEGFR-Y1068 and pEGFR-Y1173 were 80.3% (335/417) and 82.2% (347/422), respectively, indicating distinct functional roles of phosphorylated proteins from total proteins.

We then applied DeepEnrich to analyze tEGFR, pEGFR-Y1068, and pEGFR-Y1173 in LUAD (Table 1). tEGFR was significantly enriched in chr7p11, a region where *EGFR* is located, but not pEGFRs. Both tEGFR and pEGFRs were significantly enriched in the SHEDDEN_LUNG_CANCER_GOOD_SURVIVAL_A4, which contains several differentiation-related genes.[59] The leading-edge genes for tEGFR were unique from those in pEGFRs, since the pLEGs of tEGFR were excluded during the identification of phospho-specific pLEGs. Only six leading-edge genes overlapped between the two pEGFRs (*ATP11A*, *CAPN2*, *EPHX3*, *ICAM1*, *NDNF*, and *SPTB*). This suggests that while tEGFR and pEGFRs share common regulatory pathways, they have unique pLEGs that may contribute to distinct functional roles in lung cancer survival and differentiation. In addition, the three proteins were enriched in different lung epithelial cell types. Specifically, only pEGFR-Y1173 was enriched in signaling alveolar epithelial type 2 cells, reported as putative alveolar stem cells[60] and as the origin of LUAD.[61,62]

As a member of the ErbB family, HER2/ERBB2 alterations are occasionally observed in LUAD.[63] While HER2 and its phosphorylated form, pHER2-Y1248, are highly correlated in BRCA, their correlation is low in LUAD (Figure S5B). HER2 preferentially heterodimerizes with other ErbB receptors,[64,65] which may underlie our observation that pHER2 was highly correlated with pEGFRs, particularly pEGFR-Y1068 (Figure S5C). HER2, but not pHER2, was significantly enriched in Chr17q12, the locus of *ERBB2* (Table 1), whereas pHER2 was enriched in alveolar epithelial type 1 cells. Consistent with this observation, prior studies demonstrated HER2 in the differentiation of lung epithelium,[66] and its expression and activation in lung epithelium.[67–69] These results underscore cancer- and site-specific regulation of phosphoproteins. A full list of enriched pathways for pEGFR and pHER2 in LUAD is provided in Tables S12 and S13 for pLEG and nLEG, respectively.

Predicted protein abundance of DeepGxP associated with overall survival in BRCA

We performed survival analyses on all genes and proteins for estrogen receptor-positive BRCA samples ($n = 626$). A total of 15 genes and 19 proteins were significantly associated with overall survival (Figures 4A and 4B). Only Bcl-2 and cyclin B1 were significant at both gene and protein levels, underlining the divergence between RNA and protein molecular patterns. The discordance was exemplified for p53, where protein abundance, but not *TP53* mRNA expression, was significantly associated with poorer survival (Figure 4B). To validate this, we applied DeepGxP to samples lacking RPPA data ($n = 180$). Consistent with measured data, the predicted p53 protein abundance showed a significant association with poorer survival (Cox $p = 0.034$), whereas *TP53* expression alone did not (Figure 4C). Supporting this, *TP53* mutant samples showed higher p53 protein abundance than wild-type samples (Figure S6). High p53 protein abundance correlates with TP53 mutations,[70–73] which in turn are linked to poor survival,[74] elevated expression of Ki67,[75] and higher tumor grade[76] in estrogen receptor-positive breast cancer. These results demonstrate that survival cannot always be inferred from gene expression alone, as regulatory processes downstream of transcription often decouple mRNA and protein abundances. Inferring proteins with DeepGxP provided features more proximal to clinical outcomes, and several self-genes

Table 1. Enrichment of pLEGs among EGFR, pEGFRs, HER2, and pHER2

Gene set	EGFR	EGFR_pY1068	EGFR_pY1173	HER2	HER2_pY1248
	N1/N2 (*p* value)	N1/N2 (*p* value)	N1/N2 (*p* value)	N1/N2 (*p*-value)	N1/N2 (*p* value)
Prognostic					
SHEDDEN_LUNG_CANCER_GOOD_SURVIVAL_A4	**14/197 (6.53×10^{-8})**	**23/197 (3.79×10^{-14})**	**18/197 (2.38×10^{-9})**	**22/197 (8.64×10^{-14})**	**15/197 (1.07×10^{-7})**
Genomic position					
Chr7p11 (EGFR)	**11/83 (3.29×10^{-12})**	NA	1/83 (0.502)	NA	**8/83 (1.9×10^{-7})**
Chr17q12 (ERBB2)	3/144 (0.057)	**8/144 (2.53×10^{-5})**	1/144 (0.702)	**14/144 (6.36×10^{-12})**	3/144 (0.092)
Lung cell type					
TRAVAGLINI_LUNG_ALVEOLAR_EPITHELIAL_TYPE_1_CELL	**19/391 (1.27×10^{-7})**	11/391 (0.041)	11/391 (0.052)	16/391 (1.58×10^{-4})	**20/391 (6.09×10^{-7})**
TRAVAGLINI_LUNG_ALVEOLAR_EPITHELIAL_TYPE_2_CELL	4/136 (0.071)	5/136 (0.06)	**11/136 (1.26×10^{-5})**	NA	NA
TRAVAGLINI_LUNG_SIGNALING_ALVEOLAR_EPITHELIAL_TYPE_2_CELL	4/71 (0.009)	6/71 (7.95×10^{-4})	**8/71 (1.79×10^{-5})**	**7/71 (6.57×10^{-5})**	3/71 (0.076)
TRAVAGLINI_LUNG_CLUB_CELL	5/114 (0.01)	**10/114 (1.09×10^{-5})**	7/114 (0.002)	**9/114 (3.61×10^{-5})**	7/114 (1.07×10^{-3})
TRAVAGLINI_LUNG_NEUTROPHIL_CELL	10/366 (0.01)	9/366 (0.115)	**18/366 (2.87×10^{-5})**	NA	NA

N1 represents the number of pLEGs matched to the pathway gene set, and N2 represents the size of the pathway gene set. *p* values were calculated using Fisher's exact test. $p < 0.0001$ are highlighted in bold.

that were non-prognostic at the RNA level became predictive at the protein level.

Implementation of DeepGxP to single-cell profiles

Our model configuration can also be applied to single-cell data generated by cellular indexing of transcriptomes and epitopes (CITE) sequencing (CITE-seq).[77] DeepGxP was trained using a dataset of 53,364 human peripheral blood mononuclear cells (PBMCs); 19,738 genes (after filtering genes with low expression from an initial set of 20,729 genes prior to SCTransform normalization); and 224 proteins[78] (see STAR Methods). The model's architecture is depicted in Figure 5A. We modified the bulk RNA version to incorporate Markov affinity-based graph imputation of cells (MAGIC) imputation[79] after SCTransform normalization and reduced the 1DConv filter size from 1,024 to 256.

For protein abundance prediction, DeepGxP demonstrated highest performance (median, $\rho = 0.576$) compared to XGB (median, $\rho = 0.535$), sciPENN[80] (median, $\rho = 0.497$), and linear regression (median, $\rho = 0.220$) (Figure 5B). Moreover, DeepGxP outperformed self-gene prediction, with 80% (162/203) of proteins reaching a correlation coefficient above 0.31 between predicted and actual abundance (the threshold for the high-sGP group), compared to only 29% (59/203) for self-gene prediction (Figure 5C). Using CLEC12A as an example which had the best prediction performance, DeepGxP enhanced the already strong self-gene correlation coefficient to 0.97, showing a 36.6% improvement (Figure 5D).

We further validated DeepGxP using two internal datasets from time points 3 and 7 (with time point 0 as training), and external datasets from cord blood mononuclear cells (CBMCs)[77] and H1N1-influenza PBMCs.[81] Validation focused on markers for T cells and monocytes (CD4, CD8, CD14, and CD16). DeepGxP performed better for all four markers than sciPENN for CBMCs (Figure S6 and Table 2), despite DeepGxP being trained on PBMCs, which differ in blood source location and immune cell maturity. We also tested DeepGxP on our own experimental data generated using Abseq (BD Biosciences),[82] which is conceptually similar to CITE-seq. The dataset consisted of PBMCs from three healthy individuals. All three samples have at least 82.6% (19/23) of proteins achieving $\rho > 0.3$ between DeepGxP predicted and true abundance (Table S14).

DeepGxP showed high accuracy in predicting CD16 protein abundance, strongly correlated with true abundance in the H1N1 dataset (Figure S7). Natural killer (NK) cells ($median_{CD16\text{-}NK} = 0.78$) and non-classical monocytes ($median_{CD16\text{-}nCM} = 0.84$) had higher protein abundance compared to classical monocytes ($median_{CD16\text{-}CM} = 0.12$) (Figure 5E). This occurred despite low *FCGR3A* RNA expression in NK cells ($median_{FCGR3A\text{-}NK} = 0.04$) and classical monocytes ($median_{FCGR3A\text{-}CM} = 0.00$), and the fact that CD16 protein was elevated only in NK cells. Conversely, although the model predicted high CD16 abundance for NK cells and non-classical monocytes, the latter had a 5-fold higher *FCGR3A* RNA expression than NK cells ($median_{FCGR3A\text{-}nCM} = 0.20$). These results demonstrate that DeepGxP accurately predicts protein abundance independently of RNA expression levels and captures cell-type-specific variations.

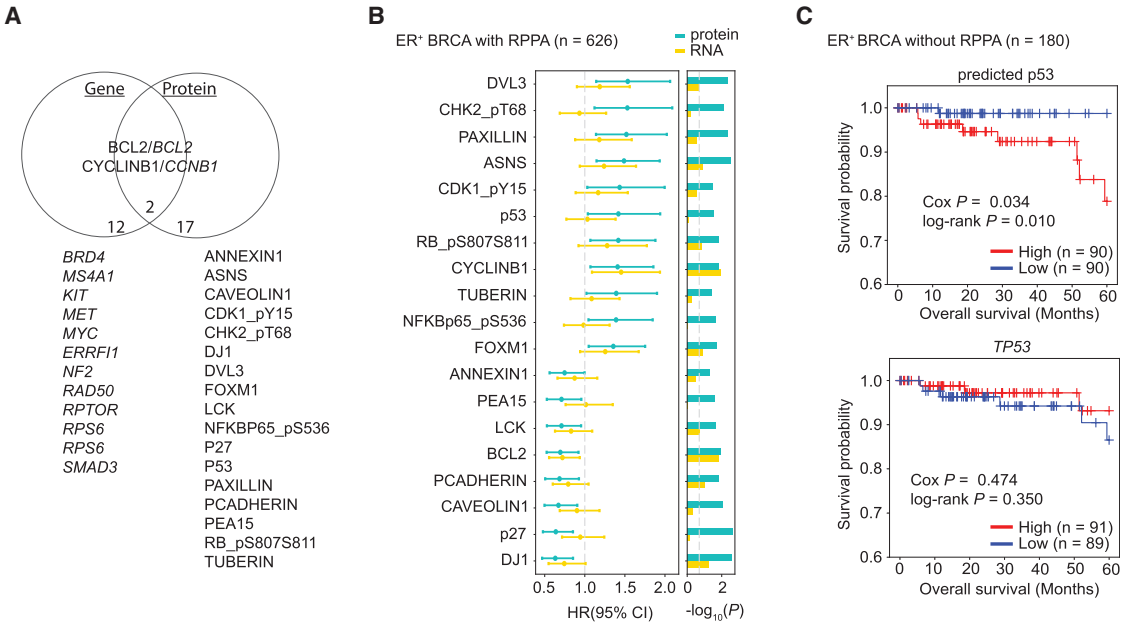


Figure 4. Validation of clinical significance of predicted survival-associated proteins on ER + BRCA samples

(A) Survival-associated ($p < 0.05$) genes and proteins using the Cox-regression model based on samples in training/validation/testing sets.

(B) Hazard ratios of survival-associated proteins. Only proteins with statistically significant differences (Cox $p < 0.05$) are shown.

(C) DeepGxP-predicted p53 abundance (top) and expression of its corresponding gene (*TP53*, bottom) using estrogen receptor-positive breast cancer samples from TCGA, without RPPA measurement. Breast cancer samples were divided by the median of RNA expression or protein abundance into high and low groups.

Louvain clustering using predicted protein abundance more effectively discriminated between classical and non-classical monocytes than RNA data alone (Figure 5F). At a resolution of 0.1, predicted protein abundance produced two distinct clusters, but only one for RNA data. Moreover, the predicted protein abundance showed better alignment with cell type labels, demonstrating the utility of predicted protein abundance in cell-type classification.

In summary, DeepGxP demonstrated significantly better prediction of protein abundance for both bulk RNA and single-cell CITE-seq data. It showed biological implications with total proteins and phosphoproteins through the DeepEnrich framework and revealed cell-type-specific protein dynamics in single-cell data, highlighting its potential for exploring complex biological processes.

DISCUSSION

This study addresses a substantial challenge in understanding the human genome—the fact that information from the transcript level can only partially represent protein abundance. We introduce DeepGxP, a deep learning model designed to uncover complex non-linear relationships between the transcriptome and proteome while providing interpretable biological insights through its associated framework, DeepEnrich. Our design is based on the concept of the translation process and considers the complex regulatory network from the transcriptome that supports the prediction of the proteome. DeepGxP employs a conventional one-dimensional CNN model, optimized through hyperparameter searching (Table S2A). Despite its simplicity, DeepGxP outperformed multiple complex deep learning models (Figure 1C), consistent with our previous work.[45] Although trained on only ~7,000 tumor samples from TCGA, DeepGxP effectively leveraged whole-transcriptome data to predict protein abundance, including phosphorylated proteins beyond the translation process. Moreover, DeepGxP's predictive advantage extends to single-cell data analyses, underscoring its broad applicability.

Functional classification of proteins revealed superior performance of DeepGxP with poor mRNA–protein concordance (Figure S3). Proteins involved in RNA metabolism (ADAR1 and RBM15), translation (eEF2 and JAB1), and DNA metabolism (ERCC5, RAD51, and XRCC1) exhibited poor mRNA-protein correlations, possibly due to their extensive post-transcriptional and post-translational regulatory networks. Conversely, calcium-binding proteins (Annexins), cell surface receptors (EGFR, HER2, HER3, and c-KIT), and scaffold/adaptor proteins (GAB2 and Caveolin-1) demonstrated stronger correlations, reflecting more direct transcriptional control of their abundance. Interestingly, DeepGxP achieved its most substantial performance gains for proteins with weak mRNA-protein correlations, suggesting that DeepGxP can effectively capture the complex regulatory mechanisms governing these challenging protein classes. This finding indicates that rather than excluding poorly correlated proteins from predictive analyses, whole-transcriptome modeling approaches can extend proteome predictability across diverse functional categories.

Cancer-type heterogeneity shapes gene-protein relationship, with protein abundance largely driven by tumor lineage, followed by pathway involvement.[39,83] DeepEnrich revealed

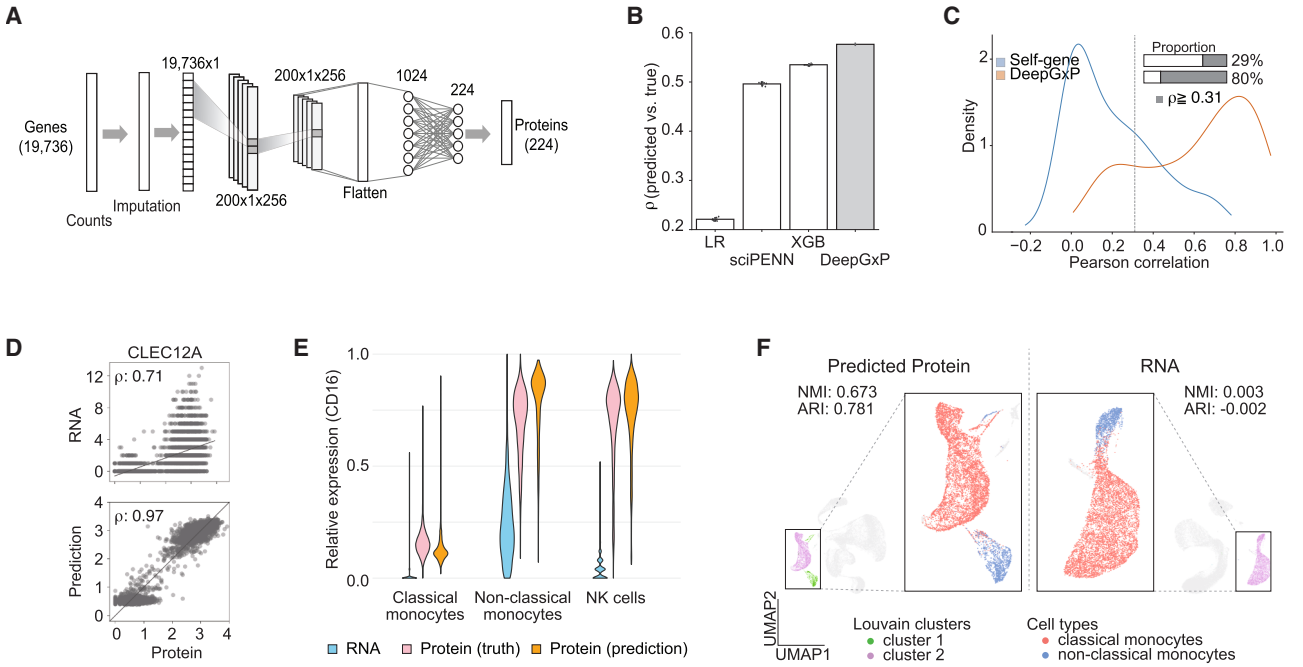


Figure 5. DeepGxP implementation in single-cell profiles

(A) Model structure of DeepGxP for single-cell applications, in which an imputation layer was added.

(B) Performance comparison of DeepGxP with XGB, sciPENN, and linear regression (LR) models. Average per-protein Pearson's correlation from 10-fold cross-validations. Each dot represents the result from the testing set after each fold training. Data shown as mean ± standard deviations.

(C) Per-protein Pearson's correlation between measured protein abundance and DeepGxP-predicted (orange) and self-gene expression (blue). The dashed line marks 0.31, the threshold for high-sGP.

(D) Prediction performance of CLEC12A. Ground truth versus self-gene expression (left) and ground truth versus DeepGxP-predicted protein abundance (right).

(E) Relative expression of CD16, a monocyte marker, comparing RNA, true protein, and predicted protein across different cell types in the H1N1 dataset.

(F) Uniform manifold approximation and projection (UMAP) plots were generated based on predicted protein abundance (n = 224) and RNA expression of highly variable genes (n = 2,000). Classical and non-classical monocytes are colored according to Louvain clusters and cell type annotations (enlarged) as described in the original publication. NMI: normalized mutual information; ARI: adjusted Rand index.

proteins such as cyclins that show consistent regulation across multiple cancer types, supporting its use in pan-cancer analyses. In contrast, proteins such as EGFR exhibit cancer-specific regulation, underscoring the importance of conducting separate analyses for different cancer types. Although not included in the main results, additional analyses of BRCA subtypes revealed that models trained within individual subtypes performed markedly worse than the pan-cancer model, primarily due to limited sample sizes. Consistent with this observation, reducing the training sample size by half led to a decrease in prediction accuracy in pan-cancer model. These observations suggest that DeepGxP benefits from information sharing across tumor types while preserving sensitivity to lineage-specific regulation and that large and diverse datasets are essential for achieving robust and generalizable protein abundance predictions.

Although mutation data were not explicitly included in the model, preliminary analyses on the prediction performance for total and phosphorylated forms of EGFR and HER2 remained comparable between mutant and wild-type cohorts (by *EGFR* and *ERBB2* mutation status) in LUAD. This suggests that the downstream effects of genomic alterations are partly captured in the transcriptome, allowing DeepGxP to infer protein abundance robustly despite mutation-related heterogeneity. These studies highlight the potential of DeepGxP for biomarker discovery and prognostic modeling.

Table 2. Comparison of prediction performance based on Pearson's correlations between true and predicted protein abundance by DeepGxP and sciPENN on internal and external datasets

			DeepGxP				sciPENN			
Datasets	Cells (n)	Proteins (n)	CD4	CD8	CD14	CD16	CD4	CD8	CD14	CD16
PBMC (time 3)	54,345	224	0.923	0.878	0.874	0.893	0.915	0.931	0.785	0.887
PBMC (time 7)	54,055	224	0.871	0.919	0.885	0.914	0.923	0.884	0.804	0.897
CBMC	8,617	13	0.901	0.604	0.594	0.763	0.772	0.409	0.543	0.167
H1N1	53,201	87	0.785	0.667	0.609	0.898	0.885	0.829	0.405	0.855

A unique feature of DeepGxP is its biological interpretability, achieved through DeepEnrich, which identifies regulatory pathways contributing to protein abundance. Our enrichment analyses (Figure 3C) align with known biological processes, exemplified by the cell-cycle regulation of cyclin B1 and cyclin E2. Through normalized integrated gradients, we identified key regulatory genes with strong biological relevance, reinforcing the model's capacity to generate mechanistically interpretable predictions.

DeepEnrich also captures post-translational regulatory mechanisms, particularly in protein phosphorylation, an essential cellular process that activates or inhibits cell signaling and protein activity and functions. Applying DeepEnrich to phosphorylated proteins elucidated pathway-level regulatory interactions governing phosphorylation dynamics. For example, DeepGxP distinguished phosphorylation events of EGFR at Y1068 and Y1173, demonstrating its ability to differentiate site-specific regulatory mechanisms. Additionally, pEGFR and pHER2 showed a strong correlation in abundance and enrichment of chromosomal positions. pEGFR is enriched in the region harboring *ERBB2*, while pHER2 is functionally enriched in regions containing *EGFR* (Table 1). This reciprocal enrichment of chromosome positions suggests cross-activation of these receptors, potentially driven by HER2 kinase domain mutations.[84] Our findings indicate that phosphorylation is regulated independently of total protein levels, as reflected by their corresponding activated functions and genomic events. The moderate correlation between total and phosphorylated forms ($\rho < 0.6$) highlights this approach's ability to minimize confounding effects from total protein abundance and focus on phosphorylation driven by alternative mechanisms.

Beyond bulk transcriptomics, analyses of single-cell profiles were also improved using our model. By clustering cells based on predicted protein abundances ($n = 224$), we effectively separated classical and non-classical monocytes, outperforming RNA-based clustering ($n = 2{,}000$), which yielded only a single cluster. This highlights the limitations of RNA-based classification and the advantage of using protein-based approaches for defining immune cell subpopulations. Our data-driven analyses show the advantage of utilizing protein data for accurate classification of monocyte subsets without relying on traditional methods such as cell type markers[85] or reference mapping,[78] because protein expression directly reflects cellular function and phenotypic differences. Our findings underscore that protein-based clustering enhances our understanding of immune cell heterogeneity, where valuable insights into their roles in immune responses and disease mechanisms could be further explored.

To model protein abundance from transcriptomic data, we used RPPA-based proteomic data from TCPA[36] instead of mass spectrometry-based CPTAC.[86] TCPA profiles ~8,000 tumor samples across 32 cancer types compared to ~1,000 samples across ~10 cancer types in CPTAC, providing greater sample diversity. While TCPA measures a smaller panel of 224 proteins (recently expanded to 447[83]), this yields a more manageable feature-to-sample ratio (~0.028:1 and ~0.056:1) compared to CPTAC's ~10,000 proteins (~10:1), reducing overfitting risk in deep learning models. RPPA offers high sensitivity for clinically relevant tumor-associated proteins and post-translational modifications when high-quality antibodies are available. However, mass spectrometry provides broader proteome coverage, improved detection of low-abundance proteins, and antibody-independent quantification. Thus, while our RPPA-based approach offered practical advantages for model training in this study, future work should leverage mass spectrometry data to extend transcriptome-to-proteome prediction.

Nonetheless, DeepGxP maintained robust predictive performance across these two platforms, with 96.3% (77/80) of proteins achieving Pearson's correlations above 0.3 against mass spectrometry-based abundance when the two measurement methods were concordant ($\rho > 0.3$) in BRCA cohort. This robustness, despite the modest overall agreement between RPPA and mass spectrometry (mean, $\rho = 0.52$),[83] underscores the model's ability to extract biologically meaningful RNA-protein relationships beyond platform-specific biases.

DeepGxP represents a significant advancement in linking transcriptomics to proteomics, enabling both accurate protein abundance predictions and biological interpretability. By integrating pathway-level insights through DeepEnrich, our model uncovers regulatory mechanisms driving protein expression and post-translational modifications. The clinical utility of DeepGxP in biomarker discovery, single-cell analyses, and cancer research underscores its potential for broad biomedical applications. Future enhancements incorporating multi-omics data and spatial proteomics will further solidify its role as a powerful tool for understanding gene-protein interactions at both the bulk and single-cell levels.

Limitations of the study

Despite its strengths, DeepGxP has limitations. First, the model relies only on gene expression as input and does not incorporate other molecular factors such as DNA methylation, microRNA, or genomic mutations, all of which can influence protein abundance. Incorporating these additional layers could further enhance predictive performance and mechanistic interpretability. Second, the model was trained on RPPA data measuring a predefined set of well-known cancer-associated proteins and phosphorylation sites, restricting predictions to these targets and limiting novel protein discovery. Nevertheless, these proteins are clinically important and often poorly predicted by mRNA alone, and DeepGxP enables accurate inference of their abundance from transcriptomic data. Third, the model was trained on bulk RNA and protein data that represent averaged cell effects, which may obscure cell-type-specific relationships. Although DeepGxP has been adapted for single-cell prediction in this manuscript, fully understanding gene-protein relationships in heterogeneous tumors requires analyses tumor samples at single-cell resolution. Finally, DeepGxP identifies associations rather than causality between RNA and protein, meaning that confounding factors, reverse regulation by proteins such as transcription factors, and feedback loops cannot be excluded. Resolving this directionality would require targeted experimental perturbations beyond the scope of this work. Nonetheless, through functional characterization of predictors, independent validation on the ICGC dataset, and cross-platform validation on mass spectrometry, as well as pathway-level analyses using DeepEnrich, we demonstrate that DeepGxP captures biologically meaningful regulatory signals rather than purely statistical associations.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Yidong Chen (cheny8@uthscsa.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing and publicly available data only. The TCGA RNA sequencing (RNA-seq) data are available at https://xenabrowser.net/datapages/and RPPA data at https://tcpaportal.org/tcpa/download.html. The CCLE data are available at https://depmap.org/portal/. The CITE-seq PBMC data were downloaded from (https://atlas.fredhutch.org/nygc/multimodal-pbmc/), CBMC data came from the Seurat tutorial (https://satijalab.org/seurat/articles/multimodal_vignette.html), and H1N1 data were accessed at https://doi.org/10.35092/yhjc.c.4753772.
- All original code has been deposited at GitHub (https://github.com/hmtsai2024/DeepGxP_manuscript) and Zenodo under https://doi.org/10.5281/zenodo.15233341[87] and is publicly available as of the date of publication.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This research is supported by the fund from the Graduate Students Study Abroad Program sponsored by the Ministry of Science and Technology (MOST), Taiwan, to H.-M.T., NCI Cancer Center Shared Resources P30CA54174 (to Y.C.); Cancer Prevention and Research Institute of Texas (RP220662 (to Y.C.); Fund for Innovation in Cancer Informatics (ICI Fund to Y.C.); RP190346 (to Y.C. and Y.H.). The funding sources had no role in the design of the study; collection, analysis, and interpretation of data, or in writing the manuscript.

AUTHOR CONTRIBUTIONS

Conceptualization, H.-M.T., T.-H.H., and Y.-C.C.; data curation, H.-M.T., Y.-C.C., and Y.H.; formal analysis, investigation, software, validation, visualization, and writing – original draft, H.-M.T. and T.-H.H.; methodology, H.-M.T., T.-H.H., Y.-C.C., and Y.H.; project administration, funding acquisition, resources, T.-H.H. and Y.-C.C.; supervision, E.Y.C. and Y.C.; writing – review and editing, Y.C.

DECLARATION OF INTERESTS

The authors declare no competing interests.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the authors used ChatGPT and Claude to enhance the readability and language of this work. After using this tool/service, the authors thoroughly reviewed and edited the content as needed and take full responsibility for the content of the published article.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2026.114815.

Received: May 19, 2025
Revised: September 16, 2025
Accepted: January 22, 2026
Published: January 27, 2026

REFERENCES

1. Perou, C.M., Sørlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., et al. (2000). Molecular portraits of human breast tumours. Nature *406*, 747–752. https://doi.org/10.1038/35021093.
2. Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., et al. (2006). Gene Expression Profiling in Breast Cancer: Understanding the Molecular Basis of Histologic Grade To Improve Prognosis. J. Natl. Cancer Inst. *98*, 262–272. https://doi.org/10.1093/jnci/djj052.
3. Ma, X.J., Salunga, R., Tuggle, J.T., Gaudet, J., Enright, E., McQuary, P., Payette, T., Pistone, M., Stecker, K., Zhang, B.M., et al. (2003). Gene expression profiles of human breast cancer progression. Proc. Natl. Acad. Sci. USA *100*, 5974–5979. https://doi.org/10.1073/pnas.0931261100.
4. Bovelstad, H.M., Nygard, S., Storvold, H.L., Aldrin, M., Borgan, O., Frigessi, A., and Lingjaerde, O.C. (2007). Predicting survival from microarray data–a comparative study. Bioinformatics *23*, 2080–2087. https://doi.org/10.1093/bioinformatics/btm305.
5. Sotiriou, C., and Pusztai, L. (2009). Gene-Expression Signatures in Breast Cancer. N. Engl. J. Med. *360*, 790–800. https://doi.org/10.1056/NEJMra0801289.
6. Reis-Filho, J.S., and Pusztai, L. (2011). Gene expression profiling in breast cancer: classification, prognostication, and prediction. Lancet *378*, 1812–1823. https://doi.org/10.1016/S0140-6736(11)61539-0.
7. Geeleher, P., Cox, N.J., and Huang, R.S. (2014). Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. Genome Biol. *15*, R47. https://doi.org/10.1186/gb-2014-15-3-r47.
8. Du, W., and Elemento, O. (2015). Cancer systems biology: embracing complexity to develop better anticancer therapeutic strategies. Oncogene *34*, 3215–3225. https://doi.org/10.1038/onc.2014.291.
9. Maleki, F., Ovens, K., Hogan, D.J., and Kusalik, A.J. (2020). Gene Set Analysis: Challenges, Opportunities, and Future Research. Front. Genet. *11*, 654. https://doi.org/10.3389/fgene.2020.00654.
10. Gry, M., Rimini, R., Strömberg, S., Asplund, A., Pontén, F., Uhlén, M., and Nilsson, P. (2009). Correlations between RNA and protein expression profiles in 23 human cell lines. BMC Genom. *10*, 365. https://doi.org/10.1186/1471-2164-10-365.
11. Akbani, R., Ng, P.K.S., Werner, H.M.J., Shahmoradgoli, M., Zhang, F., Ju, Z., Liu, W., Yang, J.Y., Yoshihara, K., Li, J., et al. (2014). A pan-cancer proteomic perspective on The Cancer Genome Atlas. Nat. Commun. *5*, 3887. https://doi.org/10.1038/ncomms4887.

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

13. Wang, D., Eraslan, B., Wieland, T., Hallström, B., Hopf, T., Zolg, D.P., Zecha, J., Asplund, A., Li, L.H., Meng, C., et al. (2019). A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. Mol. Syst. Biol. *15*, e8503. https://doi.org/10.15252/msb.20188503.

14. Zhang, Y., Chen, F., Chandrashekar, D.S., Varambally, S., and Creighton, C.J. (2022). Proteogenomic characterization of 2002 human cancers reveals pan-cancer molecular subtypes and associated pathways. Nat. Commun. *13*, 2669. https://doi.org/10.1038/s41467-022-30342-3.

15. Gillette, M.A., Satpathy, S., Cao, S., Dhanasekaran, S.M., Vasaikar, S.V., Krug, K., Petralia, F., Li, Y., Liang, W.W., Reva, B., et al. (2020). Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma. Cell *182*, 200–225.e35. https://doi.org/10.1016/j.cell.2020.06.013.

16. Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M.C., Zimmerman, L.J., Shaddox, K.F., Kim, S., et al. (2014). Proteogenomic characterization of human colon and rectal cancer. Nature *513*, 382–387. https://doi.org/10.1038/nature13438.

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., et al. (2016). Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. Cell *166*, 755–765. https://doi.org/10.1016/j.cell.2016.05.069.

18. Anderson, L., and Seilhamer, J. (1997). A comparison of selected mRNA and protein abundances in human liver. Electrophoresis *18*, 533–537. https://doi.org/10.1002/elps.1150180333.

19. Lundberg, E., Fagerberg, L., Klevebring, D., Matic, I., Geiger, T., Cox, J., Algenäs, C., Lundeberg, J., Mann, M., and Uhlen, M. (2010). Defining the transcriptome and proteome in three functionally different human cell lines. Mol. Syst. Biol. *6*, 450. https://doi.org/10.1038/msb.2010.106.

20. Nagaraj, N., Wisniewski, J.R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Pääbo, S., and Mann, M. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. Mol. Syst. Biol. *7*, 548. https://doi.org/10.1038/msb.2011.81.

21. Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337–342. https://doi.org/10.1038/nature10098.

22. Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A.M., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., et al. (2014). Mass-spectrometry-based draft of the human proteome. Nature *509*, 582–587. https://doi.org/10.1038/nature13319.

23. Fortelny, N., Overall, C.M., Pavlidis, P., and Freue, G.V.C. (2017). Can we predict protein from mRNA levels? Nature *547*, E19–E20. https://doi.org/10.1038/nature22293.

24. Wilhelm, M., Hahne, H., Savitski, M., Marx, H., Lemeer, S., Bantscheff, M., and Kuster, B. (2017). Wilhelm et al. reply. Nature *547*, E23. https://doi.org/10.1038/nature22294.

25. Buccitelli, C., and Selbach, M. (2020). mRNAs, proteins and the emerging principles of gene expression control. Nat. Rev. Genet. *21*, 630–644. https://doi.org/10.1038/s41576-020-0258-4.

26. Prabahar, A., Zamora, R., Barclay, D., Yin, J., Ramamoorthy, M., Bagheri, A., Johnson, S.A., Badylak, S., Vodovotz, Y., and Jiang, P. (2024). Unraveling the complex relationship between mRNA and protein abundances: a machine learning-based approach for imputing protein levels from RNA-seq data. NAR Genom. Bioinform. *6*, lqae019. https://doi.org/10.1093/nargab/lqae019.

27. Srivastava, H., Lippincott, M.J., Currie, J., Canfield, R., Lam, M.P.Y., and Lau, E. (2022). Protein prediction models support widespread post-transcriptional regulation of protein abundance by interacting partners. PLoS Comput. Biol. *18*, e1010702. https://doi.org/10.1371/journal.pcbi.1010702.

28. Yang, M., Petralia, F., Li, Z., Li, H., Ma, W., Song, X., Kim, S., Lee, H., Yu, H., Lee, B., et al. (2020). Community Assessment of the Predictability of Cancer Protein and Phosphoprotein Levels from Genomics and Transcriptomics. Cell Syst. *11*, 186–195.e9. https://doi.org/10.1016/j.cels.2020.06.013.

29. Xu, W., He, H., Guo, Z., and Li, W. (2022). Evaluation of machine learning models on protein level inference from prioritized RNA features. Brief. Bioinform. *23*, bbac091. https://doi.org/10.1093/bib/bbac091.

30. Cranney, C.W., and Meyer, J.G. (2024). Multi-dataset Integration and Residual Connections Improve Proteome Prediction from Transcriptomes using Deep Learning. Preprint at bioRxiv, 2024.07.08.602560. https://doi.org/10.1101/2024.07.08.602560.

31. Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F.J. (2019). Deep learning: new computational modelling techniques for genomics. Nat. Rev. Genet. *20*, 389–403. https://doi.org/10.1038/s41576-019-0122-6.

32. Matos, L.L.d., Trufelli, D.C., de Matos, M.G.L., and da Silva Pinhal, M.A. (2010). Immunohistochemistry as an important tool in biomarkers detection and clinical practice. Biomark. Insights *5*, 9–20. https://doi.org/10.4137/bmi.s2185.

33. Gutierrez, C., and Schiff, R. (2011). HER2: biology, detection, and clinical implications. Arch. Pathol. Lab Med. *135*, 55–62. https://doi.org/10.5858/2010-0454-RAR.1.

34. Carrasco-Zanini, J., Pietzner, M., Davitte, J., Surendran, P., Croteau-Chonka, D.C., Robins, C., Torralbo, A., Tomlinson, C., Grünschläger, F., Fitzpatrick, N., et al. (2024). Proteomic signatures improve risk prediction for common and rare diseases. Nat. Med. *30*, 2489–2498. https://doi.org/10.1038/s41591-024-03142-z.

35. Tasaki, S., Xu, J., Avey, D.R., Johnson, L., Petyuk, V.A., Dawe, R.J., Bennett, D.A., Wang, Y., and Gaiteri, C. (2022). Inferring protein expression changes from mRNA in Alzheimer's dementia using deep neural networks. Nat. Commun. *13*, 655. https://doi.org/10.1038/s41467-022-28280-1.

36. Li, J., Akbani, R., Zhao, W., Lu, Y., Weinstein, J.N., Mills, G.B., and Liang, H. (2017). Explore, Visualize, and Analyze Functional Cancer Proteomic Data Using the Cancer Proteome Atlas. Cancer Res. *77*, e51–e54. https://doi.org/10.1158/0008-5472.CAN-17-0369.

37. Akbani, R., Becker, K.F., Carragher, N., Goldstein, T., de Koning, L., Korf, U., Liotta, L., Mills, G.B., Nishizuka, S.S., Pawlak, M., et al. (2014). Realizing the promise of reverse phase protein arrays for clinical, translational, and basic research: a workshop report: the RPPA (Reverse Phase Protein Array) society. Mol. Cell. Proteomics *13*, 1625–1643. https://doi.org/10.1074/mcp.O113.034918.

38. Ding, L., Bailey, M.H., Porta-Pardo, E., Thorsson, V., Colaprico, A., Bertrand, D., Gibbs, D.L., Weerasinghe, A., Huang, K.-l., Tokheim, C., et al. (2018). Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. Cell *173*, 305–320.e10. https://doi.org/10.1016/j.cell.2018.03.033.

39. Hoadley, K.A., Yau, C., Hinoue, T., Wolf, D.M., Lazar, A.J., Drill, E., Shen, R., Taylor, A.M., Cherniack, A.D., Thorsson, V., et al. (2018). Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell *173*, 291–304.e6. https://doi.org/10.1016/j.cell.2018.03.022.

40. Campbell, J.D., Yau, C., Bowlby, R., Liu, Y., Brennan, K., Fan, H., Taylor, A.M., Wang, C., Walter, V., Akbani, R., et al. (2018). Genomic, Pathway Network, and Immunologic Features Distinguishing Squamous Carcinomas. Cell Rep. *23*, 194–212.e6. https://doi.org/10.1016/j.celrep.2018.03.063.

41. Preuer, K., Lewis, R.P.I., Hochreiter, S., Bender, A., Bulusu, K.C., and Klambauer, G. (2018). DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. Bioinformatics *34*, 1538–1546. https://doi.org/10.1093/bioinformatics/btx806.

42. Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods *12*, 931–934. https://doi.org/10.1038/nmeth.3547.

43. Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. *33*, 831–838. https://doi.org/10.1038/nbt.3300.

44. Chaudhary, K., Poirion, O.B., Lu, L., and Garmire, L.X. (2018). Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. Clin. Cancer Res. *24*, 1248–1259. https://doi.org/10.1158/1078-0432.CCR-17-0853.

45. Mostavi, M., Chiu, Y.C., Huang, Y., and Chen, Y. (2020). Convolutional neural network models for cancer type prediction based on gene expression. BMC Med. Genomics *13*, 44. https://doi.org/10.1186/s12920-020-0677-2.

46. Lyu, B., and Haque, A. (2018). Deep Learning Based Tumor Type Classification Using Gene Expression Data. Proc ACM Int Conf Bioinform Comput Biol Health Inform (BCB), 89–96. https://doi.org/10.1145/3233547.3233588.

47. Zeng, S., Adusumilli, T., Awan, S.Z., Immadi, M.S., Xu, D., and Joshi, T. (2025). G2PDeep-v2: a web-based deep-learning framework for phenotype prediction and biomarker discovery using multi-omics data. Biomolecules. *15*, 1673. https://doi.org/10.3390/biom15121673.

48. Zeng, S., Mao, Z., Ren, Y., Wang, D., Xu, D., and Joshi, T. (2021). G2PDeep: a web-based deep-learning framework for quantitative phenotype prediction and discovery of genomic markers. Nucleic Acids Res. *49*, W228–W236. https://doi.org/10.1093/nar/gkab407.

49. Nikolados, E.M., Wongprommoon, A., Aodha, O.M., Cambray, G., and Oyarzún, D.A. (2022). Accuracy and data efficiency in deep learning models of protein expression. Nat. Commun. *13*, 7755. https://doi.org/10.1038/s41467-022-34902-5.

50. Nikolados, E.M., and Oyarzún, D.A. (2023). Deep learning for optimization of protein expression. Curr. Opin. Biotechnol. *81*, 102941. https://doi.org/10.1016/j.copbio.2023.102941.

51. Zrimec, J., Fu, X., Muhammad, A.S., Skrekas, C., Jauniskis, V., Speicher, N.K., Börlin, C.S., Verendel, V., Chehreghani, M.H., Dubhashi, D., et al. (2022). Controlling gene expression with deep generative design of regulatory DNA. Nat. Commun. *13*, 5099. https://doi.org/10.1038/s41467-022-32818-8.

52. Zrimec, J., Börlin, C.S., Buric, F., Muhammad, A.S., Chen, R., Siewers, V., Verendel, V., Nielsen, J., Töpel, M., and Zelezniak, A. (2020). Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. Nat. Commun. *11*, 6141. https://doi.org/10.1038/s41467-020-19921-4.

53. Mi, H., Ebert, D., Muruganujan, A., Mills, C., Albou, L.P., Mushayamaha, T., and Thomas, P.D. (2021). PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. Nucleic Acids Res. *49*, D394–D403. https://doi.org/10.1093/nar/gkaa1106.

54. Shen, M., Feng, Y., Gao, C., Tao, D., Hu, J., Reed, E., Li, Q.Q., and Gong, J. (2004). Detection of Cyclin B1 Expression in G1-Phase Cancer Cell Lines and Cancer Tissues by Postsorting Western Blot Analysis. Cancer Res. *64*, 1607–1610. https://doi.org/10.1158/0008-5472.CAN-03-3321.

55. Erlandsson, F., Wählby, C., Ekholm-Reed, S., Hellström, A.C., Bengtsson, E., and Zetterberg, A. (2003). Abnormal expression pattern of cyclin E in tumour cells. Int. J. Cancer *104*, 369–375. https://doi.org/10.1002/ijc.10949.

56. Park, C.S., Kim, S.I., Lee, M.S., Youn, C.Y., Kim, D.J., Jho, E.H., and Song, W.K. (2004). Modulation of beta-catenin phosphorylation/degradation by cyclin-dependent kinase 2. J. Biol. Chem. *279*, 19592–19599. https://doi.org/10.1074/jbc.M314208200.

57. Davidson, G., and Niehrs, C. (2010). Emerging links between CDK cell cycle regulators and Wnt signaling. Trends Cell Biol. *20*, 453–460. https://doi.org/10.1016/j.tcb.2010.05.002.

58. Wirth, D., Özdemir, E., and Hristova, K. (2023). Quantification of ligand and mutation-induced bias in EGFR phosphorylation in direct response to ligand binding. Nat. Commun. *14*, 7579. https://doi.org/10.1038/s41467-023-42926-8.

59. Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma, Shedden, K., Taylor, J.M.G., Enkemann, S.A., Tsao, M.S., Yeatman, T.J., Gerald, W.L., Eschrich, S., Jurisica, I., Giordano, T.J., et al. (2008). Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nat. Med. *14*, 822–827. https://doi.org/10.1038/nm.1790.

60. Travaglini, K.J., Nabhan, A.N., Penland, L., Sinha, R., Gillich, A., Sit, R.V., Chang, S., Conley, S.D., Mori, Y., Seita, J., et al. (2020). A molecular cell atlas of the human lung from single-cell RNA sequencing. Nature *587*, 619–625. https://doi.org/10.1038/s41586-020-2922-4.

61. Wang, Z., Li, Z., Zhou, K., Wang, C., Jiang, L., Zhang, L., Yang, Y., Luo, W., Qiao, W., Wang, G., et al. (2021). Deciphering cell lineage specification of human lung adenocarcinoma with single-cell RNA sequencing. Nat. Commun. *12*, 6500. https://doi.org/10.1038/s41467-021-26770-2.

62. Han, G., Sinjab, A., Rahal, Z., Lynch, A.M., Treekitkarnmongkol, W., Liu, Y., Serrano, A.G., Feng, J., Liang, K., Khan, K., et al. (2024). An atlas of epithelial cell states and plasticity in lung adenocarcinoma. Nature *627*, 656–663. https://doi.org/10.1038/s41586-024-07113-9.

63. Hynes, N.E., and Schlange, T. (2006). Targeting ADAMS and ERBBs in lung cancer. Cancer Cell *10*, 7–11. https://doi.org/10.1016/j.ccr.2006.06.012.

64. Swanton, C., Futreal, A., and Eisen, T. (2006). Her2-targeted therapies in non-small cell lung cancer. Clin. Cancer Res. *12*, 4377s–4383s. https://doi.org/10.1158/1078-0432.CCR-06-0115.

65. Graus-Porta, D., Beerli, R.R., Daly, J.M., and Hynes, N.E. (1997). ErbB-2, the preferred heterodimerization partner of all ErbB receptors, is a mediator of lateral signaling. EMBO J. *16*, 1647–1655. https://doi.org/10.1093/emboj/16.7.1647.

66. Vermeer, P.D., Panko, L., Karp, P., Lee, J.H., and Zabner, J. (2006). Differentiation of human airway epithelia is dependent on erbB2. Am. J. Physiol. Lung Cell. Mol. Physiol. *291*, L175–L180. https://doi.org/10.1152/ajplung.00547.2005.

67. Vermeer, P.D., Einwalter, L.A., Moninger, T.O., Rokhlina, T., Kern, J.A., Zabner, J., and Welsh, M.J. (2003). Segregation of receptor and ligand regulates activation of epithelial growth factor receptor. Nature *422*, 322–326. https://doi.org/10.1038/nature01440.

68. Finigan, J.H., Faress, J.A., Wilkinson, E., Mishra, R.S., Nethery, D.E., Wyler, D., Shatat, M., Ware, L.B., Matthay, M.A., Mason, R., et al. (2011). Neuregulin-1-human epidermal receptor-2 signaling is a central regulator of pulmonary epithelial permeability and acute lung injury. J. Biol. Chem. *286*, 10660–10670. https://doi.org/10.1074/jbc.M110.208041.

69. Mishra, R., Foster, D.G., Finigan, J.H., and Kern, J.A. (2019). Interleukin-6 is required for Neuregulin-1 induced HER2 signaling in lung epithelium. Biochem. Biophys. Res. Commun. *513*, 794–799. https://doi.org/10.1016/j.bbrc.2019.04.070.

70. Soussi, T. (2011). TP53 mutations in human cancer: database reassessment and prospects for the next decade. Adv. Cancer Res. *110*, 107–139. https://doi.org/10.1016/B978-0-12-386469-7.00005-0.

71. Anderson, S.A., Bartow, B.B., Harada, S., Siegal, G.P., Wei, S., Dal Zotto, V.L., and Huang, X. (2024). p53 protein expression patterns associated with TP53 mutations in breast carcinoma. Breast Cancer Res. Treat. *207*, 213–222. https://doi.org/10.1007/s10549-024-07357-z.

72. Mendoza, R.P., Chen-Yost, H.I.H., Wanjari, P., Wang, P., Symes, E., Johnson, D.N., Reeves, W., Mueller, J., Antic, T., and Biernacka, A. (2024). Lung adenocarcinomas with isolated TP53 mutation: A comprehensive clinical, cytopathologic and molecular characterization. Cancer Med. *13*, e6873. https://doi.org/10.1002/cam4.6873.

73. Smardova, J., Liskova, K., Ravcukova, B., Malcikova, J., Hausnerova, J., Svitakova, M., Hrabalkova, R., Zlamalikova, L., Stano-Kozubik, K., Blahakova, I., et al. (2016). Complex analysis of the p53 tumor suppressor

in lung carcinoma. Oncol. Rep. *35*, 1859–1867. https://doi.org/10.3892/or.2015.4533.

74. Griffith, O.L., Spies, N.C., Anurag, M., Griffith, M., Luo, J., Tu, D., Yeo, B., Kunisaki, J., Miller, C.A., Krysiak, K., et al. (2018). The prognostic effects of somatic mutations in ER-positive breast cancer. Nat. Commun. *9*, 3476. https://doi.org/10.1038/s41467-018-05914-x.

75. Ellis, M.J., Ding, L., Shen, D., Luo, J., Suman, V.J., Wallis, J.W., Van Tine, B.A., Hoog, J., Goiffon, R.J., Goldstein, T.C., et al. (2012). Whole-genome analysis informs breast cancer response to aromatase inhibition. Nature *486*, 353–360. https://doi.org/10.1038/nature11143.

76. Pereira, B., Chin, S.-F., Rueda, O.M., Vollan, H.-K.M., Provenzano, E., Bardwell, H.A., Pugh, M., Jones, L., Russell, R., Sammut, S.-J., et al. (2016). The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. Nat. Commun. *7*, 11479. https://doi.org/10.1038/ncomms11479.

77. Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P.K., Swerdlow, H., Satija, R., and Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. Nat. Methods *14*, 865–868. https://doi.org/10.1038/nmeth.4380.

78. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., 3rd, Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573–3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.

79. van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A.J., Burdziak, C., Moon, K.R., Chaffer, C.L., Pattabiraman, D., et al. (2018). Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. Cell *174*, 716–729.e27. https://doi.org/10.1016/j.cell.2018.05.061.

80. Lakkis, J., Schroeder, A., Su, K., Lee, M.Y.Y., Bashore, A.C., Reilly, M.P., and Li, M. (2022). A multi-use deep learning method for CITE-seq and single-cell RNA-seq data integration with cell surface protein prediction and imputation. Nat. Mach. Intell. *4*, 940–952. https://doi.org/10.1038/s42256-022-00545-w.

81. Kotliarov, Y., Sparks, R., Martins, A.J., Mulè, M.P., Lu, Y., Goswami, M., Kardava, L., Banchereau, R., Pascual, V., Biancotto, A., et al. (2020). Broad immune activation underlies shared set point signatures for vaccine responsiveness in healthy individuals and disease activity in patients with lupus. Nat. Med. *26*, 618–629. https://doi.org/10.1038/s41591-020-0769-8.

82. Shahi, P., Kim, S.C., Haliburton, J.R., Gartner, Z.J., and Abate, A.R. (2017). Abseq: Ultrahigh-throughput single cell protein profiling with droplet microfluidic barcoding. Sci. Rep. *7*, 44447. https://doi.org/10.1038/srep44447.

83. Li, J., Liu, W., Mojumdar, K., Kim, H., Zhou, Z., Ju, Z., Kumar, S.V., Ng, P.K.-S., Chen, H., Davies, M.A., et al. (2024). A protein expression atlas on tissue samples and cell lines from cancer patients provides insights into tumor heterogeneity and dependencies. Nat. Cancer *5*, 1579–1595. https://doi.org/10.1038/s43018-024-00817-x.

84. Wang, S.E., Narasanna, A., Perez-Torres, M., Xiang, B., Wu, F.Y., Yang, S., Carpenter, G., Gazdar, A.F., Muthuswamy, S.K., and Arteaga, C.L. (2006). HER2 kinase domain mutation results in constitutive phosphorylation and activation of HER2 and EGFR and resistance to EGFR tyrosine kinase inhibitors. Cancer Cell *10*, 25–38. https://doi.org/10.1016/j.ccr.2006.05.023.

85. Ziegler-Heitbrock, L., Ancuta, P., Crowe, S., Dalod, M., Grau, V., Hart, D.N., Leenen, P.J.M., Liu, Y.J., MacPherson, G., Randolph, G.J., et al. (2010). Nomenclature of monocytes and dendritic cells in blood. Blood *116*, e74–e80. https://doi.org/10.1182/blood-2010-02-258558.

86. Li, Y., Dou, Y., Da Veiga Leprevost, F., Geffen, Y., Calinawan, A.P., Aguet, F., Akiyama, Y., Anand, S., Birger, C., Cao, S., et al. (2023). Proteogenomic data and resources for pan-cancer analysis. Cancer Cell *41*, 1397–1406. https://doi.org/10.1016/j.ccell.2023.06.009.

87. Tsai, H.-M., Hsiao, T.-H., Chiu, Y.-C., Huang, Y., Chuang, E.Y., and Chen, Y. (2025). Predicting and interpreting protein and phosphoprotein abundance from transcriptomes in pan-cancer and single-cells. https://doi.org/10.5281/zenodo.15233340.

88. Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15. https://doi.org/10.1186/s13059-017-1382-0.

89. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.

90. Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K.T., Montavon, G., Samek, W., Müller, K.-R., Dähne, S., and Kindermans, P.-J. (2019). iNNvestigate neural networks. Journal of machine learning research *20*, 1–8.

91. Goldman, M.J., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., Banerjee, A., Luo, Y., Rogers, D., Brooks, A.N., et al. (2020). Visualizing and interpreting cancer genomics data via the Xena platform. Nat. Biotechnol. *38*, 675–678. https://doi.org/10.1038/s41587-020-0546-8.

92. Hutter, C., and Zenklusen, J.C. (2018). The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. Cell *173*, 283–285. https://doi.org/10.1016/j.cell.2018.03.042.

93. Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinf. *12*, 323. https://doi.org/10.1186/1471-2105-12-323.

94. Li, J., Lu, Y., Akbani, R., Ju, Z., Roebuck, P.L., Liu, W., Yang, J.Y., Broom, B.M., Verhaak, R.G.W., Kane, D.W., et al. (2013). TCPA: a resource for cancer functional proteomics data. Nat. Methods *10*, 1046–1047. https://doi.org/10.1038/nmeth.2650.

95. International Cancer Genome Consortium, Hudson, T.J., Anderson, W., Artez, A., Barker, A.D., Bell, C., Bernabé, R.R., Bhan, M.K., Calvo, F., Eerola, I., et al. (2010). International network of cancer genome projects. Nature *464*, 993–998. https://doi.org/10.1038/nature08987.

96. Bolstad, B. (2022). preprocessCore: A collection of pre-processing functions (R package). Bioconductor. https://doi.org/10.18129/B9.bioc.preprocessCore.

97. He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In 2015 IEEE International Conference on Computer Vision (ICCV) (IEEE), pp. 1026–1034. https://doi.org/10.1109/ICCV.2015.123.

98. Kingma, D.P., and Ba, J. (2014). Adam: A method for stochastic optimization. Preprint at arXiv. https://doi.org/10.48550/arXiv/1412.6980.

99. Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., and Cox, D.D. (2015). Hyperopt: a Python library for model selection and hyperparameter optimization. Comput. Sci. Discov. *8*, 014008. https://doi.org/10.1088/1749-4699/8/1/014008.

100. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE), pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

101. Kingma, D.P., and Welling, M. (2013). Auto-encoding variational bayes. Preprint at arXiv. https://doi.org/10.48550/arXiv/1312.6114.

102. Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Association for Computing Machinery), pp. 785–794. https://doi.org/10.1145/2939672.2939785.

103. Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS *16*, 284–287. https://doi.org/10.1089/omi.2011.0118.

104. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation *2*, 100141. https://doi.org/10.1016/j.xinn.2021.100141.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
TCGA PANCAN RNA-seq	Pan-Cancer Atlas	https://xenabrowser.net/datapages/?cohort=TCGA%20Pan-Cancer%20(PANCAN)&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443
TCPA RPPA protein data	TCPA	https://tcpaportal.org/tcpa/download.html
CCLE mass spectrometry protein data	DepMap	https://depmap.org/portal/
CCLE RPPA protein data	DepMap	https://depmap.org/portal/
CITE-seq of healthy PBMC data	Hao et al.[78]	GEO: GSE164378
CITE-seq of CBMC	Stoeckius et al.[77]	GEO: GSE100866
CITE-seq of PBMC (H1N1)	Kotliarov et al.[81]	https://nih.figshare.com/collections/Data_and_software_code_repository_for_Broad_immune_activation_underlies_shared_set_point_signatures_for_vaccine_responsiveness_in_healthy_individuals_and_disease_activity_in_patients_with_lupus_Kotliarov_Y_Sparks_R_et_al_Nat_Med_DOI_https_d/4753772
Software and algorithms		
DeepGxP	This paper[87]	https://github.com/hmtsai2024/DeepGxP_manuscript; Zenodo: https://doi.org/10.5281/zenodo.15233341
sciPENN	Lakkis et al.[80]	https://github.com/jlakkis/sciPENN
scanpy	Wolf et al.[88]	https://github.com/scverse/scanpy
Scipy v1.5.4	Virtanen et al.[89]	https://github.com/scipy/scipy
Seurat v4	Hao et al.[78]	https://github.com/satijalab/seurat
MAGIC v2.0.3	van Dijk et al.[79]	https://github.com/KrishnaswamyLab/MAGIC
iNNvestigate v1.0.9	Alber[90]	https://github.com/albermax/innvestigate

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

This study did not involve the recruitment of new experimental models or study participants. All analyses were conducted using previously published, publicly available human datasets, including TCGA,[39] PBMC (GSE164378),[78] CBMC (GSE100866),[77] and the H1N1[81] dataset. Comprehensive information regarding participant demographics, sample sizes, inclusion and exclusion criteria, experimental design, and institutional review board (IRB) approval is available in the corresponding original publications.

No samples were excluded or reallocated based on sex, gender, ancestry, race, or ethnicity in the present study.

METHOD DETAILS

Model construction datasets

Gene expression data (11,069 samples and 20,531 genes) were downloaded from the UCSC Xena Browser under the Pan-Cancer Atlas Hub (EB++AdjustPANCAN_IlluminaHiSeq_RNASeqV2.geneExp.xena, captured in 2017).[91] The data, generated by the TCGA PanCan Atlas project,[39,92] had been preprocessed prior to release, including batch-normalization and $\log_2$ transformation (log_2 *(norm_value* + 1)), where *norm_value* represents RSEM-derived counts.[93] Genes with low expression and variation (mean <1, SD <0.6) were removed, except for those encoding the 187 retained proteins. Expression values were then z-transformed to standardize their contribution to protein prediction.

Level 4 proteome data (7,754 samples, 258 proteins) measured by reverse phase protein arrays (RPPA) were downloaded from the TCPA portal (https://tcpaportal.org/tcpa/download.html, TCGA-PANCAN32-L4).[94] These data had been batch-corrected and median-normalized prior to release, representing relative protein abundance on the $\log_2$ scale.[83] To ensure consistency, only solid tumor types with > 50 samples and paired mRNA-protein measurements were included (n = 6,813), while hematopoietic and lymphoid samples were excluded to avoid lineage differences.

Protein measurement reproducibility was assessed using cross-platform validation with data from the Cancer Cell Line Encyclopedia (CCLE), downloaded from the Broad Institute's DepMap Portal (https://depmap.org/portal/). The CCLE dataset includes protein abundance quantified by both RPPA (214 proteins, CCLE_RPPA_20181003) and mass spectrometry (12,399 proteins, protein_quant_current_normalized) across 362 cell lines. Proteins showing low reproducibility between the two platforms (Pearson correlation < 0.3 or $P > 0.05$) were excluded, resulting a final set of 187 proteins for modeling. After filtering, the final dataset comprised 13,995 genes and 187 proteins across 6,813 solid tumors from 27 tumor types (Table S1).

Model validation datasets

For external validation, RNA and protein data from the ICGC[95] were downloaded from the UCSC Xena Browser (RNA: 8,004 samples, 20,503 genes; protein: 3,242 samples, 505 proteins). RNA-seq gene expression data were normalized by read counts. A total of 3,053 samples had both RNA and protein measurements, with 169 proteins matching the protein list used in the training dataset. To ensure compatibility with the TCGA-based training data, RNA expression data from the ICGC were quantile-normalized to the distribution of the TCGA dataset using the preprocessCore R package.[96]

DeepGxP architecture and training

DeepGxP was developed using Keras 2.3.1 under the TensorFlow backend. The model architecture includes a one-dimensional convolutional layer (1024 filters, kernel length 50) followed by one flattened layer, one fully connected layer (512 nodes), and an output prediction layer (Figure 1A). Model parameters were initialized using He's uniform distribution[97] and weights were updated using the Adam optimizer.[98] Training data were split into 80%/10%/10% for training, validation, and testing, with early stopping applied after five epochs without improvement in validation loss. Hyperparameters were optimized using the Hyperas method[99] (Table S2A).

Deep and machine learning model comparison

To benchmark DeepGxP against alternative models, we implemented both deep learning and classical machine learning approaches. Deep learning methods included a modified ResNet[100] and a variational autoencoder (VAE),[101] while classical machine learning methods included XGBoost,[102] support vector machine (SVM), linear regression (LR), multivariate linear regression (mLR), and canonical correlation analysis (CCA). Model performance was evaluated using mean squared error (MSE) and Pearson correlation (ρ) in 10-fold cross-validation.

Implementation of comparative models

All methods were configured for regression tasks and took the entire gene expression matrix (13,995 genes) as input, except for univariate linear regression, which used only self-gene expression (defined as the transcript that codes for the corresponding protein). Machine learning models were trained and tested using functions from the XGBoost (version 1.5.2) and Scikit-learn (version 0.24.2) Python libraries and trained with default hyperparameters. DeepGxP, ResNet, VAE, and CCA were trained to predict multiple proteins simultaneously, whereas XGBoost and the other three machine learning models were trained to predict single proteins per model.

ResNet model

The original ResNet[100] architecture was adapted for regression by converting its two-dimensional convolutional layers to one-dimensional and replacing the final classification layer with a linear activation output layer. Several ResNet versions were evaluated (v1: 8, 14, 20 layers; v2: 11, 20, 29 layers), with ResNet v1 (8 layers) showing the best performance.

Variational autoencoder (VAE)

Unlike a conventional autoencoder, which maps the input data into a fixed latent vector, the VAE[101] learns to encode the inputs into a distribution $N(\mu, \sigma^2)$ for each element of the latent vector. Its loss function was the average of the reconstruction error and Kullback-Leibler (KL) divergence. The reconstruction error is computed as the mean square error between true and predicted outputs. The KL divergence regularizes the learned latent distribution by minimizing the divergence between the learned distribution and a Gaussian distribution N(0, 1). Hyperparameters, including the numbers of hidden layers, fully connected layers, and nodes, were selected using Hyperas[99] (Table S2B).

Model baseline

To confirm that model performance was not driven by random correlations, Y-scrambling was performed in which the output proteins were randomly permuted while inputs genes remained unaltered. This analysis served as a baseline to verify that DeepGxP's predictive performance was from true biological signal rather than chance and random correlation.

Minimal sample size requirement for predicting protein abundance

To assess the learning capability of DeepGxP to unseen tumor types, DeepGxP was retrained without breast cancer samples. A set of testing samples (n=87) was withheld from training for performance evaluation. Breast cancer samples were gradually added to the training set with sample sizes of 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, and 698. For each addition, the model was trained repetitively 10 times. To evaluate the minimum sample size at which DeepGxP stopped learning robustly, we calculated the incremental rate of the performance metric as the difference in performance between consecutive sample sizes (e.g., analogously between time points A, B, C, and D). The turning point was identified when the growth rate between two consecutive time points became negative, i.e. when the difference in performance between consecutive sample sizes (B-A, C-B, or D-C) was negative—indicating a decrease or diminishment in performance improvement. To refine the curve and ensure smooth trend visualization, the

locally weighted scatterplot smoothing (Lowess) method was applied to the performance curve. This approach allowed us to pinpoint the optimal sample size for including breast cancer samples and assess the model's ability to generalize across unseen tumor types.

Model interpretation via DeepEnrich

DeepEnrich was developed to discover factors associated with each protein prediction and to annotate their biological functions and pathways. To identify genes contributing significantly to predictions of protein abundance, we computed integrated gradients (IG) using the iNNvestigate Python package (version 1.0.9). IG evaluates the output gradient with respect to small changes in input features, calculating the path of gradients from a baseline expression (set to 0) to the actual input expression.

The contribution of the i-th gene to the prediction of the j-th protein, given an input instance of n dimension (number of genes), $\boldsymbol{x} \in \mathcal{R}^n$, is computed using the IG method:

$$IG_{ij}(x)::=\left(x_i-x_i'\right)\times\int_{a=0}^{1}\frac{\partial F_j(\boldsymbol{x}'+a(\boldsymbol{x}-\boldsymbol{x}'))}{\partial x_i}da \tag{1}$$

where x_i' is the baseline expression of gene i; x_i is the actual expression of gene i; F_j is the model function for protein j,and a is the scaling coefficient.

The contribution of each gene to protein j abundance prediction was quantified by averaging IG values across the top 10% of samples with the highest protein abundance (S_{Top}) to capture the most relevant gene-protein relationships:

$$aIG_{i,j}=\frac{1}{|S_{Top}|}\sum\nolimits_{s\in S_{Top}}IG_{i,j,s} \tag{2}$$

where $aIG_{i,j}$ is the average IG for the i-th gene of the j-th protein, and S_{Top} is the set of top 10% samples ranked by protein j abundance.

The averaged IG values were z-transformed to obtain normalized IG (nIG) across all genes:

$$nIG_{i,j}=\frac{aIG_{i,j}-\mu_{aIG_j}}{\sigma_{aIG_j}} \tag{3}$$

where $nIG_{i,j}$ is the z-transformed average IG for the i-th gene of j-th protein; μ_{aIG_j} is the mean aIG across all genes; and σ_{aIG_j} is the standard deviation of aIG across all genes.

For each protein j, genes with nIG values exceeding ±1.96 ($P<0.05$, assuming nIG possesses normal distribution) were defined as positive or negative leading edge (pLEG or nLEG) predictors, respectively, representing genes that significantly contribute to the prediction of protein j abundance:

$$LEG_j=\begin{cases}pLEG, if\ nIG_{i,j}>1.96\\ nLEG, if\ nIG_{i,j}<-1.96\\ not\ significant, otherwise\end{cases} \tag{4}$$

To assess whether pLEG and nLEG predictors selected by DeepGxP were enriched in specific biological functions, we conducted an over-representation analysis using the clusterProfiler[103,104] package in R. Enrichment analyses were performed separately for pLEG and nLEG predictors across multiple gene sets, including Gene Ontology (GO), KEGG, Reactome, Hallmark, and MSigDB collections (C1, C3, C4, C5, C6, C7, C8). For the GO database, redundant terms were filtered based on kappa statistics. Terms with significant kappa p-values ($p<0.05$) were clustered, and the most representative term for each cluster was selected by the lowest enrichment p-value and largest gene set size.

Definition of phospho-specific predictors

For phosphoprotein analyses, phospho-specific predictors were identified by removing predictors shared with their corresponding total proteins. Specifically, for each phosphoprotein p_i and its paired total protein t_i, the sets of pLEG and nLEG were defined as:

$$\begin{aligned}\text{phospho}-\text{pLEG}\ (p_i)&=\text{pLEG}_{\text{phospho}}(p_i)\setminus\text{pLEG}_{\text{total}}(t_i)\\ \text{phospho}-\text{nLEG}\ (p_i)&=\text{nLEG}_{\text{phospho}}(p_i)\setminus\text{nLEG}_{\text{total}}(t_i)\end{aligned} \tag{5}$$

where p_i represents the phosphoprotein, t_i represents the paired total protein, and $\text{pLEG}_{\text{phospho}}$ and $\text{pLEG}_{\text{total}}$ represent the sets of pLEG predictors for phosphoprotein and total protein, respectively (Figure S7). This approach allows us to identify distinguish genes whose contributions are unique to phosphorylation events.

Definition and selection of self-genes

For proteins encoded by multiple transcripts, the one with the highest Pearson correlation to protein abundance was selected. Based on the Pearson correlation between protein abundance and self-gene expression across all 187 proteins, proteins were classified into

two groups: those with a correlation above the median (0.31) were deemed the high-correlation group (high-sGP), while those below the median were deemed the low-correlation group (low-sGP).

Survival analyses

A univariate Cox proportional hazards regression model was implemented to analyze overall associations between survival, gene expression, and protein abundance in estrogen receptor-positive breast cancer samples. Kaplan-Meier curves were calculated using median values as cutoffs to define high and low gene expression and protein abundance, and *p*-values were calculated using the Cox model.

DeepGxP implementation in single-cells

Training and internal validation

To deploy DeepGxP for single-cell profiles, we trained and validated the model using the largest publicly available CITE-seq (Cellular Indexing of Transcriptomes and Epitopes by sequencing) dataset in terms of protein number: human peripheral blood mononuclear cells (PBMCs) comprising 161,764 cells, 20,953 genes, and 224 proteins (GSE164378).[78] The dataset, generated using 3′ prime (3P) RNA libraries and antibody-derived tag (ADT) panels, was accessed from the New York Genome Center (https://atlas.fredhutch.org/nygc/multimodal-pbmc/). Data preprocessing, including quality control and filtering, was performed as described in the original publication. Specifically, the dataset had excluded cells containing fewer than 500 genes, more than 6,000 detected genes, fewer than 50,000 ADT reads, or fewer than 10,000 hashtag oligo (HTO) reads.

For model training, DeepGxP was trained on timepoint 0 cells (53,364 cells), while timepoints 3 (54,345 cells) and 7 (54,055 cells) were used for internal validation. Genes for training were filtered based on the timepoint 0 data, retaining those expressed in at least 3 cells and with at least 200 detected features, resulting in 19,736 genes prior to normalization.

For external validation, we used two independent CITE-seq datasets: the cord blood mononuclear cells (CBMCs) dataset (8,617 cells, 20,501 genes, and 13 proteins; GSE100866),[77] accessed from the Seurat tutorial (https://satijalab.org/seurat/articles/multimodal_vignette.html), and H1N1-influenza PBMCs (H1N1) dataset (53,201 cells, 51,735 genes, and 87 proteins),[81] accessed via https://doi.org/10.35092/yhjc.c.4753772.

Data normalization and imputation

Gene expression data were normalized using SCTransform, and protein abundance data were normalized using the centered log ratio transformation within each cell, as implemented in Seurat (v4.2.0). After normalization, data were imputed using Markov affinity-based graph imputation of cells (MAGIC). For external validation datasets, missing genes were filled with zeros after MAGIC imputation.

Model design

We modified two key designs to optimize DeepGxP's performance for CITE-seq data. First, the MAGIC algorithm (v2.0.3)[79] was applied to the count data after SCTransform normalization, addressing excess zeros. Second, the convolutional filter size of the Conv1D layer was reduced from 1024 to 256 to better capture sparse single-cell data patterns. These adjustments were tailored to mitigate zero inflation in single-cell datasets and enhance protein abundance prediction.

Comparison to another deep learning method

To benchmark DeepGxP's performance, we compared it to one of the best deep learning models available, sciPENN,[80] a model based on a series of feedforward blocks and recurrent neural networks. We trained sciPENN with the same input and output data as DeepGxP, using default settings.

Clustering and UMAP in CITE-seq data

Clustering of predicted protein and RNA data were performed using the Louvain method with a resolution of 0.1 via the FindClusters function in Seurat for the H1N1 dataset. For clustering, we used 224 proteins and the 2000 most variable genes. Uniform manifold approximation and projection was applied to the clustering results to reduce the data to two dimensions for visualization. Cell annotations were assigned based on labels provided in the original publication.[81] We evaluated consistency between the cluster assignments and the original cell type annotations using the Adjusted Rand Index and Normalized Mutual Information.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analyses were performed using Python scipy (version 1.5.4). All data were checked for normality by the Shapiro test. T-tests were used for normally distributed data, while Mann-Whitney or Wilcoxon signed-rank tests were used for otherwise nonparametric distributed data, as appropriate. No multiple comparisons were performed in each test.