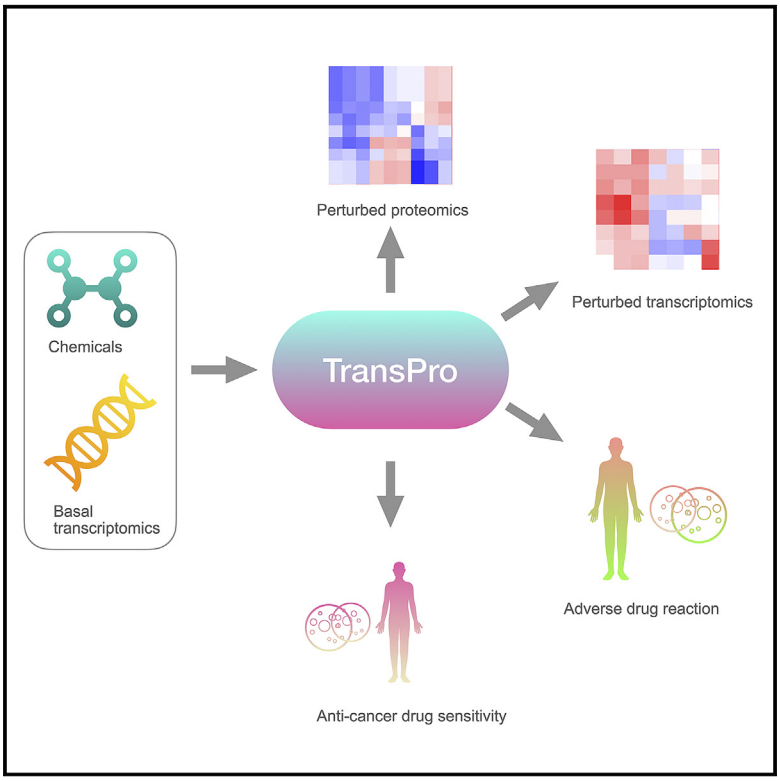
Cell Reports
Methods

Article

Hierarchical multi-omics data integration and modeling predict cell-specific chemical proteomics and drug responses

Graphical abstract



Authors

You Wu, Qiao Liu, Lei Xie

Correspondence

lxie@iscb.org

In brief

Wu et al. present TransPro, an end-to-end deep-learning framework that predicts chemical-induced, cell line-specific proteomics profiles and cellular and organismal phenotypes from transcriptomics data through hierarchical integration of multiomics data. TransPro enables systems pharmacology-driven compound screening for complex diseases.

Highlights

- TransPro is a method for predicting chemical-induced, cell line-specific proteomics
- An end-to-end deep learning model predicts drug potencies and side effects
- A multi-omics data integration strategy simulates biological information transmission
- TransPro enables systems pharmacology-oriented compound screening

Wu et al., 2023, Cell Reports Methods 3, 100452
April 24, 2023
https://doi.org/10.1016/j.crmeth.2023.100452

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Article

Hierarchical multi-omics data integration and modeling predict cell-specific chemical proteomics and drug responses

You Wu,[1] Qiao Liu,[1] and Lei Xie[1,2,3,4,*]
[1]The Graduate Center, City University of New York, New York, NY 10016, USA
[2]Hunter College, City University of New York, New York, NY 10065, USA
[3]Weill Cornell Medicine, Cornell University, New York, NY 10021, USA
[4]Lead contact
*Correspondence: lxie@iscb.org
https://doi.org/10.1016/j.crmeth.2023.100452

MOTIVATION Transcriptomics is powerful for compound screening in systems pharmacology but may not capture a comprehensive picture of biological processes. Proteomics, on the other hand, may provide a more complete view of molecular function by measuring changes in protein expression and post-translational modifications, but it is expensive, and limited proteomics coverage introduces the problem of significant missing data. To address these problems, we developed TransPro, a deep learning model that predicts chemical proteomics profiles for uncharacterized cell lines using transcriptomics data and explicitly models the information transmission from RNAs to proteins.

SUMMARY

Drug-induced phenotypes result from biomolecular interactions across various levels of a biological system. Characterization of pharmacological actions therefore requires integration of multi-omics data. Proteomics profiles, which may more directly reflect disease mechanisms and biomarkers than transcriptomics, have not been widely exploited due to data scarcity and frequent missing values. A computational method for inferring drug-induced proteome patterns would therefore enable progress in systems pharmacology. To predict the proteome profiles and corresponding phenotypes of an uncharacterized cell or tissue type that has been disturbed by an uncharacterized chemical, we developed an end-to-end deep learning framework: TransPro. TransPro hierarchically integrated multi-omics data, in line with the central dogma of molecular biology. Our in-depth assessments of TransPro's predictions of anti-cancer drug sensitivity and drug adverse reactions reveal that TransPro's accuracy is on par with that of experimental data. Hence, TransPro may facilitate the imputation of proteomics data and compound screening in systems pharmacology.

INTRODUCTION

Given the high cost and low success rate of the conventional drug discovery process, systems pharmacology has emerged as a new drug discovery paradigm.[1,2] Several recent studies have demonstrated the potential of systems pharmacology in tackling complex diseases such as Alzheimer's disease[3,4] and cancers.[5] Chemical-induced omics profiling (e.g., transcriptomics) is a potentially powerful assay readout for systems pharmacology-oriented compound screening[6,7] because it can provide an unbiased assessment of the drug's therapeutic effect on disease molecular phenotypes and critical information on the drug mode of actions. A great deal of effort has been devoted to collecting, annotating, and predicting chemical-induced transcriptomics profiles.[8–10] However, RNA expression alone may not capture a comprehensive picture of biological processes. Because molecular functions mainly manifest at a protein level, a proteomics profile may be better at characterizing and predicting cellular and organismal phenotypes than a transcriptomics profile.[11] For example, the AVIL gene has been identified as a bona fida oncogene for glionoma.[12] Although there is a significant difference in the protein expression of AVIL and its interacting genes between patients with glioma and controls, there is no detectable difference in the RNA expression of these genes.

Additionally, the dysregulation of post-translational modifications (e.g., phosphorylation and epigenetics) of proteins is a common molecular etiology of many diseases.[13,14] Such molecular events cannot be easily detected and are poorly correlated

with RNA expression. The experimental approach to proteomics is more expensive and time consuming than that of transcriptomics. As a result, few chemical-perturbed proteomics data are available.[11] Furthermore, there are significant missing values in the proteomics data.[15] Thus, a machine learning method for predicting chemical-perturbed proteomics will overcome the technical limitations of proteomics experiments, thereby offering new opportunities for systems pharmacology-driven drug discovery and precision medicine.

The machine learning prediction of chemical-perturbed cell-specific proteomics is challenging due to the scarcity of proteomics data. Integration of abundant transcriptomics data is a natural solution to impute missing proteomics data and predict unseen proteomics profiles of a novel cell line that neither has a measured proteomics profile nor is similar to cells with characterized proteomics when perturbed by novel chemicals that have different chemical structures from those already tested in a cell line model. Many methods have been developed to integrate multiple heterogeneous omics data from diverse sources.[16] However, most state-of-the-art techniques are unidirectional and horizontal. They ignore underlying biological relationships between omics datasets that reflect the hierarchical organization of a biological system. Ideally, the integrated genomics, transcriptomics, and proteomics data should represent the information flow from DNAs to RNAs to proteins.[17]

We have developed a deep learning model, TransPro, to address the challenges in the predictive modeling of cell-specific chemical proteomics for systems pharmacology. TransPro predicts cell line proteomics profiles after chemical perturbation when only unperturbed transcriptomic data are available. TransPro utilizes the transcriptomics data and transfers the knowledge learned from gene expressions to protein expressions following the central dogma of molecular biology. Specifically, we propose a hierarchical multi-omics integration approach that explicitly models the information transmission from RNAs to proteins. In the rigorous benchmark studies, TransPro significantly outperforms all baseline models. Furthermore, when using the predicted proteomics profiles by TransPro as features to predict cellular phenotypes of drug sensitivity and organism-level adverse drug reactions, it is more accurate than using experimental transcriptomics and proteomics data, suggesting that TransPro is a valuable tool for real-world applications.

RESULTS

Overview of TransPro

The proposed TransPro is an end-to-end multi-task deep learning model that learns a generalizable representation of proteomics perturbations and their downstream effects (drug potency and toxicity). In this study, we will refer to ''novel cell lines'' that do not have a measured proteomics profile available and ''novel chemicals,'' which do not have similar chemical structures to those that have been tested in a cell line model. We formulated the problem for predicting chemical-induced proteomics profiles as a multi-output regression task, the problem of predicting side effects as a multi-output classification task, and the problem of predicting drug response (IC50) as a regression task, respectively. The inputs of TransPro include basal transcriptomics expressions as cell line features and chemical structural information. As illustrated in Figure 1, TransPro firstly compresses the transcriptomics profile into an embedding vector. We expect that the information contained in the latent space of transcriptomics can be further translated into the latent space of proteomics in a low-dimensional space through a neural network, termed a transmitter. TransPro uses a graph neural network (GNN) network to acquire chemical embeddings for the purpose of extracting chemical structural features, which is a more expressive method of chemical representation learning according to recent research,[18–20] followed by a module for generating chemical-specific difference vectors that represent the difference between basal cell transcriptomics embedding and perturbed transcriptomics embedding induced by the chemical. A multi-head attention network is used to merge the embeddings of chemicals and cells and simulate chemical-gene interactions.[21] A transmitter later transfers the transcriptomics hidden state to the proteomics hidden state. Finally, a domain-specific decoder extracts the chemically induced hidden state from the embedding vector and encodes it in a low-dimensional space for both perturbed proteomics prediction and transcriptomics prediction. To predict adverse drug reactions, a multiple-layer perceptron (MLP) classifier is concatenated to the pre-trained proteomics hidden state. Similarly, another MLP regressor is concatenated to the pre-trained proteomics hidden state to predict anti-cancer drug sensitivity. See STAR Methods for details.

We assessed the performance of the TransPro model under three scenarios as shown in Figure 1B: (1) an in-distribution (ID) setting by random splits, (2) an out-of-distribution (OOD) setting for cell lines, and (3) an OOD setting for new drugs. In the ID setting of the random split, there were similar cell lines and drugs in the training/validation data to unseen testing data. This is a trivial situation. In contrast, the setting of OOD cell lines and OOD drugs means that the cell lines or drugs in the training set were significantly different from those unseen data in the validation/testing set. They would evaluate the performance of TransPro under the real-world scenario with the distribution shift. It notes that all measured protein expressions in a cell line perturbed by a drug were considered unseen in the testing and validation data. It is different from a conventional imputation approach, which only masks a portion of proteins in a perturbed proteomics profile as unseen for testing or validation and uses the remaining ones for training.

TransPro outperforms baseline models and other chemical feature representations

We compared TransPro with several baseline models, including k-nearest neighbor (k-NN), random forest, and vanilla neural networks. Furthermore, we assessed how well the GNN chemical representation in the original TransPro performed compared with alternative chemical feature representations including neural fingerprints[22] and ECFPs (extended-connectivity fingerprints)[23] when other components remained unchanged. Note that we reported the performance of ECFP6 as the main result, and the performance of ECFP4 can be found in Table S3. There is no significant difference between the performance of ECFP6 and that of ECFP4.

CellPress
OPEN ACCESS

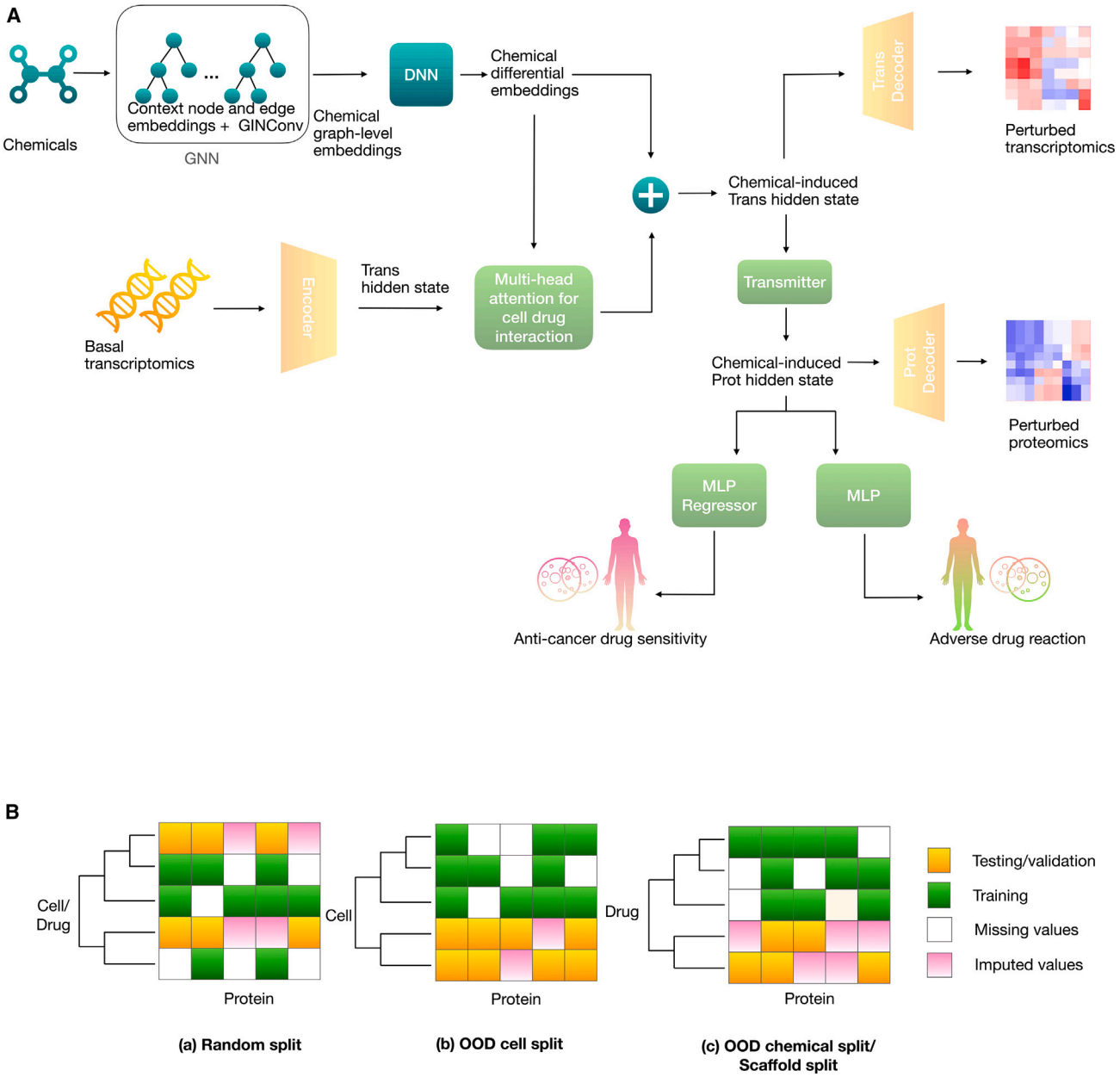


Figure 1. TransPro architecture and its performance evaluation

(A) TransPro model consists of seven major components: (1) a cell line encoder compresses basal transcriptomics to a low-dimensional vector. (2) A graph neural network (GNN) extracts the chemical embeddings. (3) A DNN for chemical-specific differential embedding generation. (4) A multi-head attention module to learn the interaction between the cell line features and the chemical features. (5) Two domain-specific decoders for transcriptomics and proteomics perturbation predictions, respectively. (6) A transmitter transfers the chemical-induced cell line features from transcriptomics latent space to proteomics latent space. (7) Task-specific classifiers/regressors for drug-induced phenotypes. Trans, Transcriptomics; Prot, proteomics.

(B) Three scenarios to evaluate TransPro performances. (1) In-distribution (ID) by the random split, where the testing data may be similar to the training and validation data. (2) Out-of-distribution (OOD) cell split, where the cell lines in testing data are significantly different from those in the training and validation data. (3) OOD chemical split, where the chemical structures in the testing data are significantly different from those in training and validation data. In the testing and validation data, no proteins in each proteomics profile are used for the training. Missing values in the testing and validation data are predicted.

TransPro outperforms baseline models and other chemical feature representations in the ID setting

We applied 3-fold cross-validation on the random split data to evaluate TransPro. Roughly 450 samples were in each fold. The result is shown in Figure 2A. TransPro outperforms all the baseline models when evaluated by both Pearson correlation and Spearman's correlation, with Pearson correlations of 8.4%, 8.4%, 12.5%, 21.5%, and 32.6% higher than ECFPs,

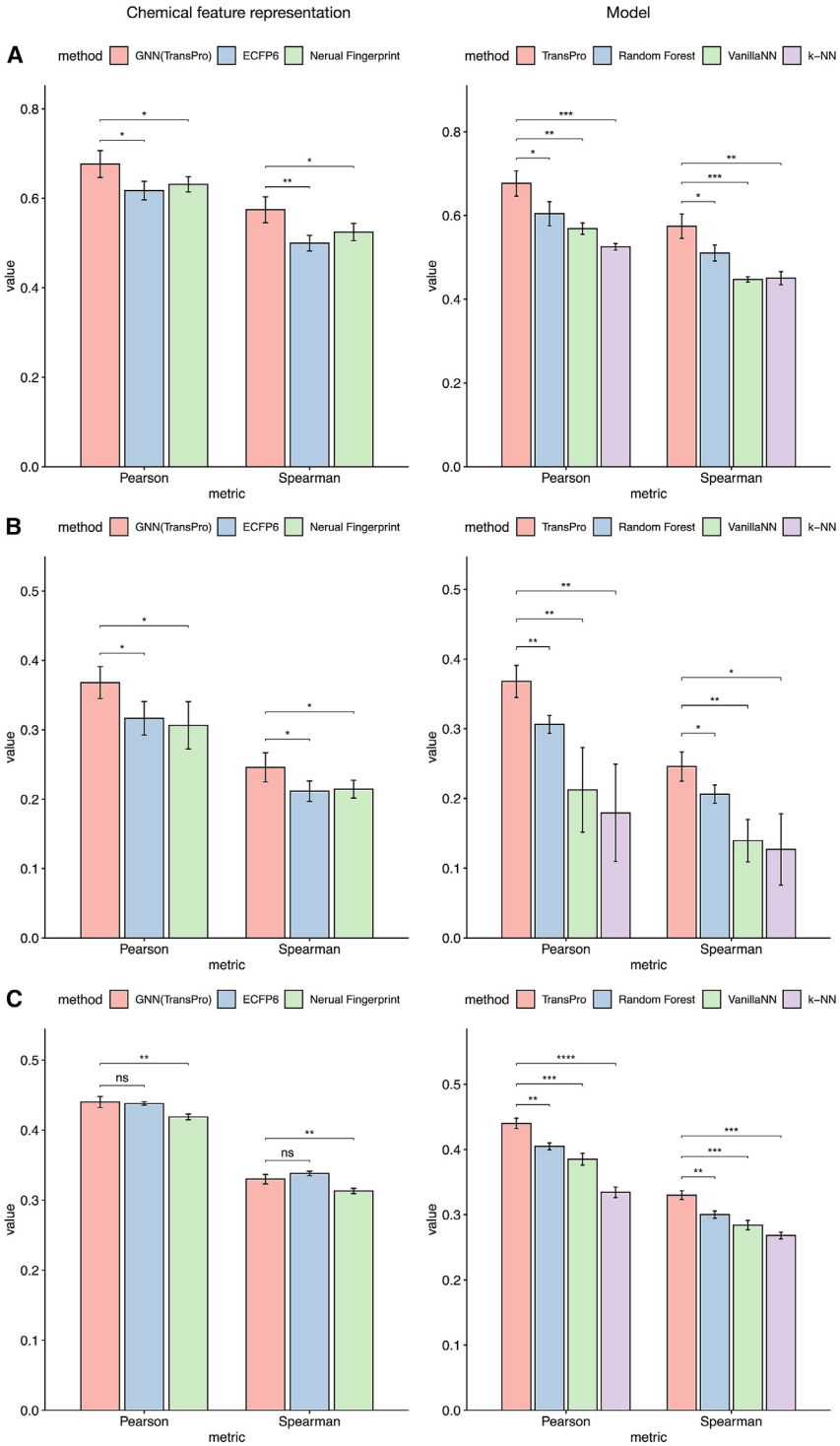


Figure 2. Performance comparison of TransPro with baseline models

(A–C) Comparisons of TransPro with different chemical feature representation methods (left panel) and baseline models (right panel) in the (A) ID setting, (B) OOD cell setting, and (C) OOD chemical setting. The stars flag levels of statistical significance. If a p value is less than 0.05, 0.01, and 0.001, it is flagged with one star (*), 2 stars (**), and three stars (***), respectively. The error bar in the figure denotes the standard deviation.

neural fingerprint, random forest, vanilla NN, and k-NN, respectively.

TransPro outperforms baseline models and other chemical feature representations in the OOD cell line setting

In order to deploy an OOD cell line setting, we ensured that the model was tested on the new cell line data that were significantly different from the cell lines in the training/validation data. We applied t-distributed stochastic neighbor embedding (t-SNE) on the cell line feature represented by the transcriptomics gene expression profile and clustered them into 3 folds manually. The cluster information is detailed in Figure S1. We applied leave-one-cluster-out 3-fold cross-validation to assess the OOD generalization abilities of TransPro for the new cell lines. The result is represented in Figure 2B. TransPro significantly outperforms all baseline models in both Pearson correlation and Spearman's correlation. When evaluated by the Pearson correlation, the performance gains over ECFPs, neural fingerprint, random forest, vanilla NN, and k-NN are 17.9%, 26.8%, 22.6%, 128.5%, and 216.5%, respectively. It is not surprising that the performance in the OOD setting of cell lines was worse than that in the ID setting of the random split, as shown in Figure 2. It is common that a machine learning model works well if training and test sets of data have the same distribution, but the performance will drop under the distribution shift.

TransPro outperforms baseline models and other chemical feature representations in the OOD chemical setting

A chemical scaffold split approach was used to split the benchmark datasets for the OOD chemical setting experiment.[24] Measuring OOD generalization is critical in molecular representation learning, where distributional shifts are enormous and challenging to control for machine learning models. RDKit[25] is used to capture the Murcko scaffold of each molecule, and only compounds with the same scaffold are grouped together. Randomly permuted groups are added to the training, validation, and testing sets. This technique assures that the testing set contains only compounds with scaffolds that are distinct from those used in the

training and validation sets. As a result, the scaffold split enables a more precise assessment of the model's ability to predict proteomics perturbations by structurally distinct chemicals. As shown in Figure 2C, TransPro consistently outperforms most of the baseline models when evaluated by both Pearson correlation and Spearman's correlation, with Pearson correlations of up to 5.9%, 9.3%, 14.9%, and 32.5% higher than neural fingerprint, random forest, vanilla NN, and k-NN, respectively. ECFPs achieved relatively equivalent performance with TransPro on the OOD chemical setting. Similar to the OOD cell line setting, the performance in the OOD chemical setting dropped compared with the random split setting due to the distribution shift. More sophisticated techniques such as chemical structure pre-training and semi-supervised learning[26,27] are needed to address the OOD challenge in the chemical space. Additionally, note that one of the differences between GNN and ECFP chemical representations is that GNN is fine-tuned but ECFP is fixed during the downstream supervised training. It has been found that the fine-tuning may distort the pre-trained features in a standard pre-training fine-tuning strategy used in this article and may deteriorate the OOD performance.[28] The performance of TransPro in the OOD chemical setting could be improved using different training procedures, e.g., one proposed by Kumar et al.[28]

In summary, TransPro outperformed or was comparable to all baseline models in the OOD settings for both OOD cell lines and OOD chemicals. Although the available transcriptomics and proteomics data for the training were not large, the generalization power of TransPro is reasonable, as suggested by training curves in which the performance of testing was slightly better than that of validation data (Figure S3).

Predicted proteomics profile has strong predictive power for adverse drug reactions

Chemical transcriptomics profiles from the LINCS1000 are effective in predicting organismal phenotypes such as adverse drug reactions.[29] Recent studies have shown that chemical-induced proteomics is more informative than chemical transcriptomics for elucidating a drug's mode of action and predicting cellular drug responses.[11] However, experimental proteomics data contain a significant number of missing values and are presumed to be suboptimal in terms of their predictive capacity for the downstream task. To explore the potential power of chemical-induced proteomics predictions, we conducted experiments to see if our predicted proteomics profile was more effective than experimental transcriptomics data and experimental proteomics data for predicting adverse drug reactions (ADRs).

ADRs are classified using the Medical Dictionary for Regulatory Activities (MedDRA) v.16.047's preferred terms (PTs). We gathered data from two ADR datasets: the off-label FDA Adverse Event Reporting System (FAERS)[30] and an on-label ADRs side-effect resource (SIDER).[31] We used a multi-label cross-entropy loss function to construct a deep neural network (DNN) as the classifier for this downstream task. For the first experiments, we examined the ADR prediction performance when using the experimentally determined LINCS1000 level-5 data and the predicted perturbed proteomics profile as features. We also clustered the LINCS1000 data into low-confidence and high-confidence data based on the Pearson correlation score between bioduplicates. Samples having a score greater than 0.5 are considered to be of high confidence; otherwise, they are low confidence. As demonstrated in the previous section, we utilized a chemical scaffold split to evaluate performance. We used the macro average for both the AUROC (area under the curve score-receiver operating characteristic) and AUPRC (area under the curve score-precision recall) as the evaluation metrics. The result is presented in Figures 3A and 3B. TransPro embedding consistently outperforms experimental transcriptomics on both metrics across two different datasets. In the FAERS low-confidence dataset, the AUROC of TransPro embedding is 6% to 6.8% higher than the experimental data, and the AUPRC is 8.7% to 15.1% higher. In the FAERS high-confidence dataset, the AUROC of TransPro embedding improves 7% to 7.1% over the experimental transcriptomics, and AUPRC improves 26.3% to 31.8%. Similar trends were observed on the SIDER dataset:'' the AUROC and AUPRC TransPro embeddings are significantly superior to the experimental transcriptomics in both high-confidence and low-confidence datasets.

In the second iteration of our study, to provide a more comprehensive evaluation, we incorporated the Cancer Perturbed Proteomics Atlas (CPPA) dataset[11] as the experimental proteomics data into our analysis. However, this dataset was characterized by a significant proportion of missing values, as shown in Figure S2. To address this issue, we employed imputation by the mean value for these missing data points. Subsequent to this process, we compared the predictive performance using the experimental proteomics data with that obtained using TransPro embeddings and experimental transcriptomics. To facilitate a rigorous comparison, we employed the intersection drugs of the LINCS1000 and CPPA datasets as our test drugs. Given that various cell lines may be responsive to a given drug, we selected the cell-drug pairs that had the highest evaluation score for each tested drug. The final score was then calculated as the mean of all tested drugs. To ensure the consistency of our training data, we filtered transcriptomics data to match the number of proteomics data points by selecting the most similar compounds to those in the experimental proteomics dataset. The results are shown in Figures 3C and 3D. The overall performance is substantially suboptimal compared with the results in Figures 3A and 3B, but TransPro embeddings still significantly outperform both experimental transcriptomics and proteomics. The inferior performance of the FAERS dataset to that of SIDER data was potentially due to the reduced sample size resulting from the intersection of experimental proteomics and transcriptomics data. The poor performance of proteomics data in all experiments is mainly due to missing values. These results suggest that a standard imputation technique is not adequate for the downstream task of chemical proteomics data. The proposed TransPro is robust for the chemical proteomics imputation and harnesses the power of chemical proteomics for drug discovery.

Predicted proteomics profile has strong predictive power for anti-cancer drug sensitivity prediction

In addition to ADRs, another valuable readout for phenotype compound screening is drug sensitivity. To explore the

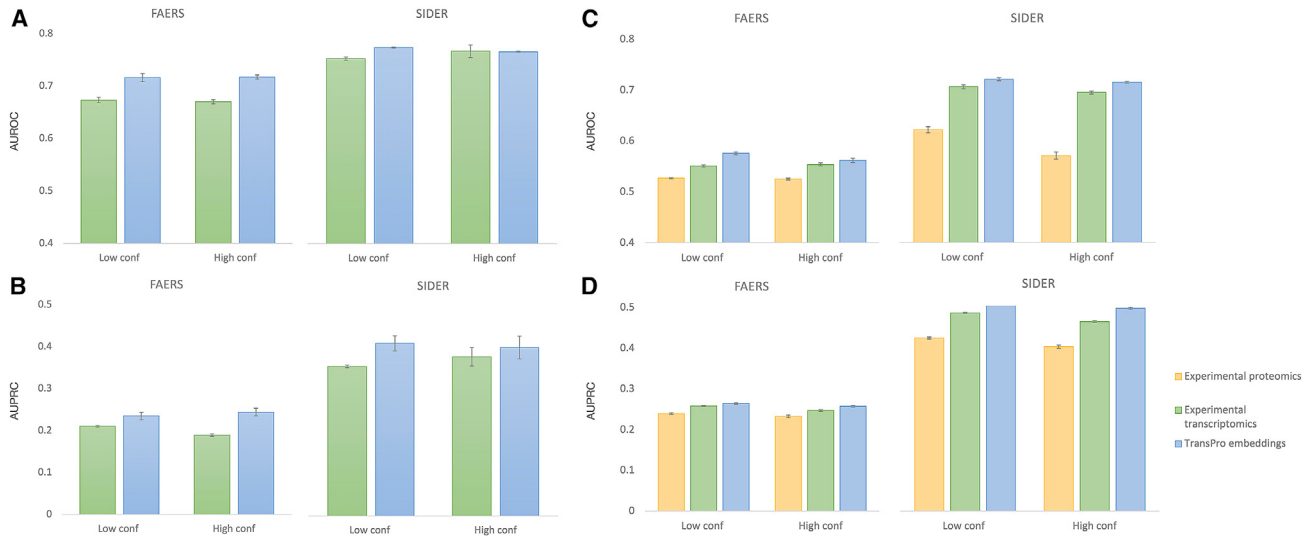


Figure 3. Model performance on the task of adverse drug reaction prediction

(A) AUROC of the side-effect prediction with all the drugs available in the experimental transcriptomics data.
(B) AUPRC of the side-effect prediction with all the drugs available in the experimental transcriptomics data.
(C) AUROC of the side-effect prediction with the shared drugs between the experimental transcriptomics and proteomics data.
(D) AUPRC of the side-effect prediction with the shared drugs between the experimental transcriptomics and proteomics data.
The error bar in the figure denotes the standard deviation.

relationship between genomic biomarkers and various drug responses, a number of large-scale genomics datasets have been generated with public access, for example Cancer Cell Line Encyclopedia (CCLE)[32] and Genomics of Drug Sensitivity in Cancer (GDSC).[33,34] The measurement of the drug sensitivity is recorded as an IC_{50} value. CCLE profiled the basal gene expressions of 1,305 cancer cell lines using RNA sequencing (RNA-seq). We used the basal gene expression profiles of cell lines in CCLE as features to predict the proteomics profiles and transcriptomics profiles first and then used the learned representations of these profiles to predict the drug sensitivity. An end-to-end training pipeline was built on top of TransPro architecture for the transcriptomics and proteomics predictions by using the drug sensitivity information to fine-tune the pre-trained TransPro model. Technically, an additional DNN head was added for this downstream task to either the transcriptomics embedding or the proteomics embedding for evaluating the predictive power of the predicted transcriptomics profiles or of the predictive proteomics profiles, respectively.

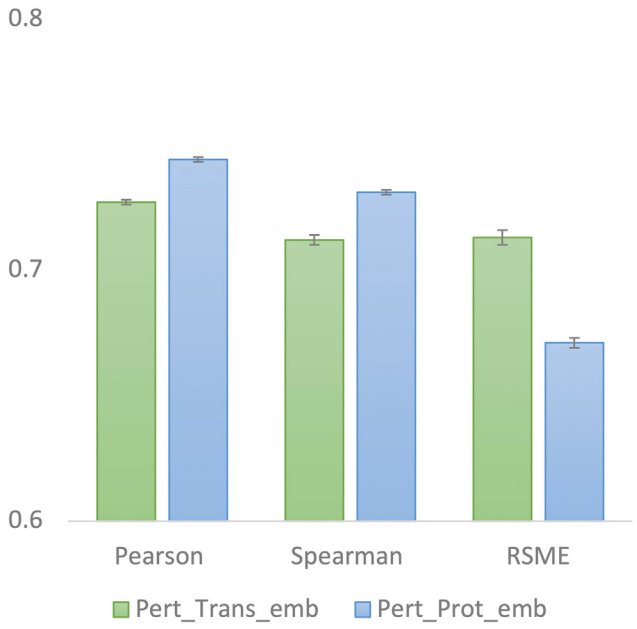


Figure 4. Comparison of drug sensitivity prediction using the predictive proteomics profiles (Pert_Trans_emb) and the predictive transcriptomics profiles (Pert_Prot_emb)

The error bar in the figure denotes the standard deviation.

As shown in Figure 4, the perturbed proteomics embedding outperforms the perturbed transcriptomics embedding by 2.3% on the Pearson correlation, 2.5% to 2.8% on the Spearman's correlation, and 5.8% to 6% on root-mean-square error (RMSE), respectively. This supports the hypothesis that the proteomics profile may have stronger predictive power for anti-cancer drug sensitivity than the transcriptomics profile.[11]

TransPro has the potential to extract drug-target information

We further evaluated if the TransPro-predicted chemical proteomics profile can detect the drug-target signals. The drug-target information was extracted from ChEMBL26.[35] The 838 drugs with a single target were chosen, and the drugs were clustered by the shared targets. The targets with multiple drugs were retained, yielding a total of 172 clusters and 656 drug-target pairs. When we used the Tanimoto score to measure in-cluster pairwise similarity between drugs, we found that 148 out of 172

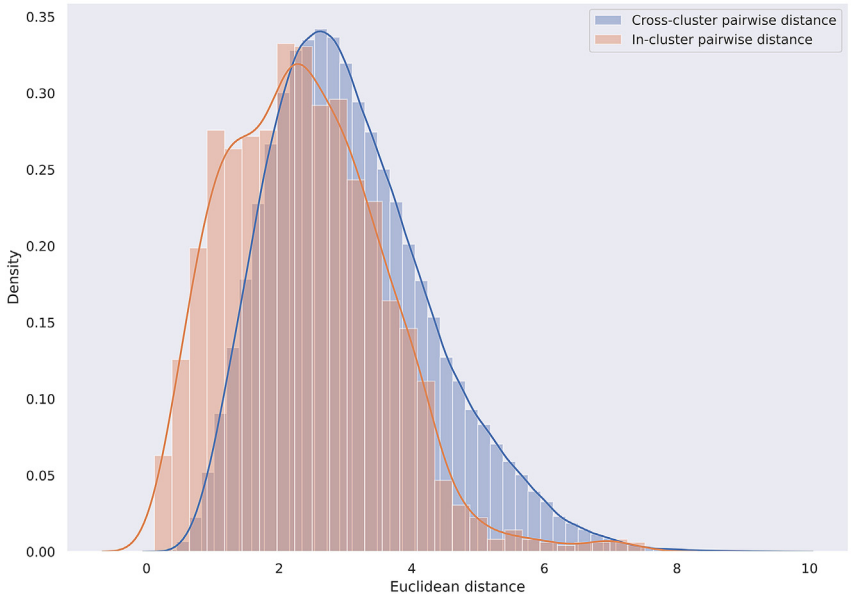


Figure 5. Distribution comparison of in-cluster and cross-cluster pairwise distances

In-cluster pairwise distance: the pairwise similarity distribution of the modeled proteomics perturbation from the drugs with the same target; cross-cluster pairwise distance: the pairwise similarity distribution of the modeled proteomics perturbation from the drugs with the different targets.

clusters had a score below 0.5, indicating that the majority of the compounds were structurally distinct, even having the same target. We paired each drug with a breast cancer cell line MCF-7 to obtain the predictive proteomics perturbation by the trained model. Noted that all the drugs from this task were OOD drugs that were not included in the training data. We calculated Euclidean distances between the predicted proteomics profiles of two drugs and generated two distance distributions: one for drug pairs within the same cluster (in cluster) and another for pairs across the cluster (cross-cluster), as shown in Figure 5. The in-cluster drugs shared the same target; thus, their proteomics profiles should be similar. As expected, the distance distribution of in-cluster drug pairs was shifted to the left, closer to zero than that of cross-cluster drug pairs that have a different target. The two distributions were significantly different as determined by the Kolmogorov-Smirnov test ($p = 7.886e{-}76$). This result suggests that the predicted proteomics profile from TransPro is biologically meaningful.

Attention module and integration of transcriptomics data contribute to the generalization of TransPro

Attention mechanisms, in which an element of one set selectively focuses on a subset of another set (cross-attention) or on its own set (self-attention), are widely used in neural network-based models and have been successfully applied to a variety of artificial intelligence tasks, including computer vision and natural language processing (NLP). In this article, we propose using the multi-head attention technique to quantify interactions between cell and chemical features. Multi-head attention was proposed for the first time in the transformer model, which delivers state-of-the-art performance on a variety of NLP tasks.[21] To determine the role of attention, we performed an ablation study on TransPro without the attention module and instead of using concatenations between cell and chemical features. The result is shown as TransPro w/o attention in Table 1 tested on three settings as aforementioned. In the ID setting, there is no significant difference between the model with attention and that without attention. However, in both OOD chemical and OOD cell settings, the use of attention improves the performance of TransPro. Thus, the attention contributes to the OOD generalization of TransPro because it allows the model to selectively focus on relevant parts of the input rather than treating all of the input equally.

Similarly, the integration of perturbed transcriptomics improved the performance of TransPro in the OOD settings compared with using proteomics alone. As shown in Table 1, by removing the perturbed transcriptomics data from the training of TransPro, the performance of the resulting model (TransPro w/o perturbed transcriptomics [PertTrans]) significantly dropped.

Applying both the attention module and PertTrans data together offers a considerable improvement over using either of them alone in the OOD cell line setting. When all components of the model are combined, the Pearson correlation improves by 2.1% to 3.8%, and the Spearman's correlation improves by 2.8%. Improvement also occurs in OOD cell settings, as the Pearson correlation increases by 12.1% to 19.1%, and the Spearman's correlation increases by 6.6% to 19.3%. Although adding neither of the two components improves performance in the ID setting, it is worth noting that the task is less challenging than in OOD chemical and OOD cell settings, and even the model without the attention module and transcriptomics data can achieve promising outcomes for the Pearson correlation up to 0.714 and the Spearman's correlation up to 0.611.

DISCUSSION

In this article, we have developed a new computational platform, TransPro, that hierarchically integrates multi-omics data for systems pharmacology-oriented compound screening. To our knowledge, TransPro is the first deep learning model for predicting cell-specific chemical proteomics profiles perturbed by unseen chemicals. Our benchmark studies have demonstrated that the performance of TransPro is acceptable for real-world applications in the OOD cell line and OOD chemical settings. Thus, TransPro can be a potentially powerful tool for systems pharmacology-driven phenotype screening. The proposed biology-inspired multi-omics data integration framework can be extended to integrate additional levels (e.g.,

Table 1. Ablation study of TransPro

Setting	Method	Pearson	Spearman	RMSE
OOD chemical	TransPro w/o Attention and PertTrans[a]	0.440 ± 0.004	0.333 ± 0.006	0.792 ± 0.020
	TransPro w/o attention	0.450 ± 0.004	0.342 ± 0.004	0.772 ± 0.013
	TransPro w/o PertTrans[a]	0.440 ± 0.008	0.330 ± 0.007	0.777 ± 0.003
	TransPro	0.460 ± 0.004	0.341 ± 0.004	0.754 ± 0.002
	p value	0.001	0.049	0.041
OOD cell	TransPro w/o attention and PertTrans[a]	0.328 ± 0.029	0.224 ± 0.021	0.723 ± 0.004
	TransPro w/o attention	0.346 ± 0.022	0.230 ± 0.024	0.684 ± 0.027
	TransPro w/o PertTrans	0.368 ± 0.023	0.246 ± 0.021	0.684 ± 0.029
	TransPro	0.378 ± 0.022	0.258 ± 0.017	0.688 ± 0.029
	p value	0.039	0.045	0.053
ID chemical and cell	TransPro w/o attention and PertTrans	0.681 ± 0.033	0.576 ± 0.035	0.593 ± 0.043
	TransPro w/o attention	0.672 ± 0.034	0.568 ± 0.035	0.600 ± 0.05
	TransPro w/o PertTrans	0.677 ± 0.030	0.574 ± 0.029	0.599 ± 0.043
	TransPro	0.674 ± 0.025	0.568 ± 0.022	0.603 ± 0.045
	p value	0.380	0.382	0.389

The p value stands for the p value of the t test between TransPro and (TransPro w/o attention and PertTrans). PertTrans, perturbed transcriptomics.
[a]The value of the t test on the evaluation metrics between TransPro and its ablated model is less than 0.05.

phosphoproteomics[36]) in a biological system for modeling genotype-phenotype associations.

Deep learning is a power horse for the success of TransPro. Firstly, because it is often infeasible to perform a large-scale compound screening with multi-omics readouts, computational prediction is necessary. Secondly, the capability of end-to-end training by deep learning makes it a powerful tool to model multi-level information transmission (e.g., from DNA to RNA to protein) and hierarchy organizations in biology (e.g., Gene Ontology).[17] Finally, advanced deep learning techniques make it possible to integrate unlabeled, heterogeneous, biased, noisy, and sparse omics data generated from diverse resources.[37]

Limitations of the study

Despite its promising results, there are several limitations of the study that might be taken into consideration.

One limitation of TransPro is the "black box" nature of deep learning. While deep learning methods have shown great potential for solving complex problems, they lack mechanistic insight into the problem learned. Although methods exist to assess feature importance,[38] new methods are needed to interpret the embeddings in the intermediate layer of a neural network, particularly with regard to proteomics features in TransPro.

Another limitation of TransPro is the challenge of training an optimal model using multiple diverse data via multiple stages for balancing the performance in both ID and OOD settings. It remains an unsolved problem on the best pre-training fine-tuning strategy for a specific problem.[39] This is an important area for future research, as it could help to improve the generalizability of deep learning models in a range of applications.

Finally, while TransPro attempts to model the underlying information transmission in a biological system, it is only in a rudimentary form. There is potential for improvement by encoding more complex information such as context-dependent gene-gene interactions and integrating additional omics data (e.g., DNA methylations) in a deep learning system.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Data
  - Baseline models
  - TransPro modules
  - Model training
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Performance evaluation

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2023.100452.

ACKNOWLEDGMENTS

This work has been supported by the National Institute of General Medical Sciences of the National Institutes of Health (R01GM122845) to L.X. and the National Institute on Aging of the National Institutes of Health (R01AD057555) to L.X.

AUTHOR CONTRIBUTIONS

Y.W. conceived the concept, prepared data, implemented the algorithms, performed the experiments, analyzed data, and wrote the manuscript; Q.L.

conceived the concept, prepared and analyzed data, and wrote the manuscript; and L.X. conceived the concept, planned the experiments, and wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 22, 2022
Revised: December 28, 2022
Accepted: March 22, 2023
Published: April 17, 2023

REFERENCES

1. Xie, L., Draizen, E.J., and Bourne, P.E. (2017). Harnessing big data for systems pharmacology. Annu. Rev. Pharmacol. Toxicol. *57*, 245–262.
2. Danhof, M. (2016). Systems pharmacology–towards the modeling of network interactions. Eur. J. Pharm. Sci. *94*, 4–14.
3. Taubes, A., Nova, P., Zalocusky, K.A., Kosti, I., Bicak, M., Zilberter, M.Y., Hao, Y., Yoon, S.Y., Oskotsky, T., Pineda, S., et al. (2021). Experimental and real-world evidence supporting the computational repurposing of bumetanide for apoe4-related alzheimer's disease. Nat. Aging *1*, 932–947.
4. Sayed, F.A., Kodama, L., Fan, L., Carling, G.K., Udeochu, J.C., Le, D., Li, Q., Zhou, L., Wong, M.Y., Horowitz, R., et al. (2021). Ad-linked r47h-trem2 mutation induces disease-enhancing microglial states via akt hyperactivation. Sci. Transl. Med. *13*, eabe3947.
5. Misek, S.A., Newbury, P.A., Chekalin, E., Paithankar, S., Doseff, A.I., Chen, B., Gallo, K.A., and Neubig, R.R. (2022). Ibrutinib blocks yap1 activation and reverses braf inhibitor resistance in melanoma cells. Mol. Pharmacol. *101*, 1–12.
6. Tan, R.K., Liu, Y., and Xie, L. (2022). Reinforcement learning for systems pharmacology-oriented and personalized drug design. Expert Opin. Drug Discov. *17*, 849–863. (just-accepted).
7. Vincent, F., Nueda, A., Lee, J., Schenone, M., Prunotto, M., and Mercola, M. (2022). Phenotypic drug discovery: recent successes, lessons learned and new directions. Nat. Rev. Drug Discov. *21*, 899–914.
8. Pham, T.-H., Qiu, Y., Zeng, J., Xie, L., and Zhang, P. (2021). 03) A deep learning framework for high-throughput mechanism-driven phenotype compound screening and its application to covid-19 drug repurposing. Nat. Mach. Intell. *3*, 1–11.
9. Wu, Y., Liu, Q., Qiu, Y., and Xie, L. (2022). Deep learning prediction of chemical-induced dose-dependent and context-specific multiplex phenotype responses and its application to personalized alzheimer's disease drug repurposing. PLoS Comput. Biol. *18*, e1010367.
10. Pham, T.-H., Qiu, Y., Liu, J., Zimmer, S., O'Neill, E., Xie, L., and Zhang, P. (2022). Chemical-induced gene expression ranking and its application to pancreatic cancer drug repurposing. Patterns (N Y) *3*, 100441.
11. Zhao, W., Li, J., Chen, M.J.M., Luo, Y., Ju, Z., Nesser, N.K., Johnson-Camacho, K., Boniface, C.T., Lawrence, Y., Pande, N.T., et al. (2020). Large-scale characterization of drug responses of clinically relevant proteins in cancer cell lines. Cancer Cell *38*, 829–843.e4.
12. Xie, Z., Janczyk, P., Zhang, Y., Liu, A., Shi, X., Singh, S., Facemire, L., Kubow, K., Li, Z., Jia, Y., et al. (2020). A cytoskeleton regulator avil drives tumorigenesis in glioblastoma. Nat. Commun. *11*, 3457.
13. Kannaiyan, R., and Mahadevan, D. (2018). A comprehensive review of protein kinase inhibitors for cancer therapy. Expert Rev. Anticancer Ther. *18*, 1249–1270.
14. Kelly, T.K., De Carvalho, D.D., and Jones, P.A. (2010). Epigenetic modifications as therapeutic targets. Nat. Biotechnol. *28*, 1069–1078.
15. Jin, L., Bi, Y., Hu, C., Qu, J., Shen, S., Wang, X., and Tian, Y. (2021). A comparative study of evaluating missing value imputation methods in label-free proteomics. Sci. Rep. *11*, 1760.
16. Lee, B., Zhang, S., Poleksic, A., and Xie, L. (2019). Heterogeneous multi-layered network model for omics data integration and analysis. Front. Genet. *10*, 1381.
17. He, D., and Xie, L. (2021). A cross-level information transmission network for hierarchical omics data integration and phenotype prediction from a new genotype. Bioinformatics *38*, 204–210.
18. Xu, K., Hu, W., Leskovec, J., and Jegelka, S. (2018). How powerful are graph neural networks?*10*.
19. Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., and Sun, M. (2020). Graph neural networks: a review of methods and applications. AI Open *1*, 57–81.
20. Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., and Leskovec, J. (2019). Strategies for pre-training graph neural networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1905.12265.
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., and Polosukhin, I. (2017). Attention Is All You Need. Advances in neural information processing systems, 30.
22. Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., and Adams, R. (2015). Convolutional networks on graphs for learning molecular fingerprints. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'15) (MIT Press), pp. 2224–2232.
23. Rogers, D., and Hahn, M. (2010). Extended-connectivity fingerprints. J. Chem. Inf. Model. *50*, 742–754.
24. Pei, J., and Zhavoronkov, A. (2021). Artificial intelligence for drug discovery and development. In Frontiers Research Topics (Frontiers Media SA), pp. 167–225.
25. Landrum, G. (2016). Rdkit: Open-source cheminformatics software, *149*, p. 650, http://www.rdkit.org/, https://github.com/rdkit/rdkit.
26. Liu, Y., Lim, H., and Xie, L. (2022). Exploration of chemical space with partial labeled noisy student self-training and self-supervised graph embedding. BMC Bioinf. *23*, 1–21.
27. Liu, Y., Wu, Y., Shen, X., and Xie, L. (2021). Covid-19 multi-targeted drug repurposing using few-shot learning. Front. Bioinform. *1*, 693177.
28. Kumar, A., Raghunathan, A., Jones, R., Ma, T., and Liang, P. (2022). Fine-tuning can distort pretrained features and underperform out-of-distribution. Preprint at arXiv. https://doi.org/10.48550/arXiv.2202.10054.
29. Wang, Z., Clark, N.R., and Ma'ayan, A. (2016). Drug-induced adverse events prediction with the lincs l1000 data. Bioinformatics *32*, 2338–2345.
30. Tatonetti, N.P., Ye, P.P., Daneshjou, R., and Altman, R.B. (2012). Data-driven prediction of drug effects and interactions. Sci. Transl. Med. *4*, 125ra31.
31. Kuhn, M., Letunic, I., Jensen, L.J., and Bork, P. (2016). The sider database of drugs and side effects. Nucleic Acids Res. *44*, D1075–D1079.
32. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., et al. (2012). 03) the cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603–607.
33. Yang, W., Soares, J., Greninger, P., Edelman, E.J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J.A., Thompson, I.R., et al. (2013). Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. *41*, D955–D961.
34. Iorio, F., Knijnenburg, T.A., Vis, D.J., Bignell, G.R., Menden, M.P., Schubert, M., Aben, N., Gonçalves, E., Barthorpe, S., Lightfoot, H., et al. (2016). A landscape of pharmacogenomic interactions in cancer. Cell *166*, 740–754.
35. Gaulton, A., Hersey, A., Nowotka, M., Bento, A.P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L.J., Cibrián-Uhalte, E., et al. (2017). The chembl database in 2017. Nucleic Acids Res. *45*, D945–D954.
36. Dele-Oni, D.O., Christianson, K.E., Egri, S.B., Vaca Jacome, A.S., DeRuff, K.C., Mullahoo, J., Sharma, V., Davison, D., Ko, T., Bula, M., et al. (2021).

Proteomic profiling dataset of chemical perturbations in multiple biological backgrounds. Sci. Data *8*, 226.

37. He, D., Liu, Q., Wu, Y., and Xie, L. (2022). A context-aware deconfounding autoencoder for robust prediction of personalized clinical drug response from cell-line compound screening. Nat. Mach. Intell. *4*, 879–892.
38. Liu, Q., and Xie, L. (2021). Transynergy: mechanism-driven interpretable deep neural network for the synergistic prediction and pathway deconvolution of drug combinations. PLoS Comput. Biol. *17*, e1008653.
39. Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., Zhang, L., et al. (2021). Pre-trained models: past, present and future. AI Open *2*, 225–250.
40. Keenan, A.B., Jenkins, S.L., Jagodnik, K.M., Koplev, S., He, E., Torre, D., Wang, Z., Dohlman, A.B., Silverstein, M.C., Lachmann, A., et al. (2018). The library of integrated network-based cellular signatures nih program: system-level cataloging of human cells response to perturbations. Cell Syst. *6*, 24.
41. Wu, Y. (2023a). Transpro datasets v2.0. https://doi.org/10.5281/zenodo.7699298.
42. Wu, Y. (2023b). Adorableyoyo/transpro_a: v1.0. https://doi.org/10.5281/zenodo.7729440.
43. Qiu, Y., Lu, T., Lim, H., and Xie, L. (2020). A Bayesian approach to accurate and robust signature detection on LINCS L1000 data. Bioinformatics *36*, 2787–2795.
44. Fescharek, R., Kübler, J., Elsasser, U., Frank, M., and Güthlein, P. (2004). Medical dictionary for regulatory activities (meddra). Int. J. Pharmaceut. Med. *18*, 259–269.
45. Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., et al. (2021). Pubchem in 2021: new data content and improved web interfaces. Nucleic Acids Res. *49*, D1388–D1395.
46. Willighagen, E.L., Mayfield, J.W., Alvarsson, J., Berg, A., Carlsson, L., Jeliazkova, N., Kuhn, S., Pluskal, T., Rojas-Chertó, M., Spjuth, O., et al. (2017). The chemistry development kit (cdk) v2.0: atom typing, depiction, molecular formulas, and substructure searching. J. Cheminform. *9*, 53.
47. Davis, J., and Goadrich, M. (2006). In The relationship between precision-recall and ROC curves (ICML '06. ACM Press), pp. 233–240.
48. Saito, T., and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. PLoS One *10*, e0118432.
49. Fawcett, T. (2004). Roc graphs: notes and practical considerations for researchers. Mach. Learn. *31*, 1–38.

CellPress
OPEN ACCESS

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
CPPA	Zhao et al.[11]	https://tcpaportal.org/cppa/#/download
LINCS	Qiu et al.[40]	https://github.com/njpipeorgan/L1000-bayesian
CCLE	Barretina et al.[32]	https://depmap.org/portal/download
Adverse Drug Reaction	Wang et al.[29]	http://maayanlab.net/SEP-L1000/#download
Drug Sensitivity	Yang et al.,[33] Iorio et al.[34]	https://www.cancerrxgene.org/downloads/bulk_download
Software and algorithms		
TransPro v1.0	This paper	https://doi.org/10.5281/zenodo.7729440

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Lei Xie (lxie@iscb.org).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The data used in this study can be accessed at https://doi.org/10.5281/zenodo.7699298.[41]
- The source code can be accessed at https://github.com/AdorableYoyo/TransPro_a.[42] DOIs are provided in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Data

Perturbed proteomics data

Through https://tcpaportal.org/cppa/#/download, [11] perturbed proteomics data were downloaded from the CPPA. The experiments involve the use of 168 different drugs and 319 distinct cell lines. Quantitative proteomics analysis was performed using reverse-phase protein arrays (RPPA). Each experiment has the incubation of cells with or without drug perturbation. Proteomics data were gathered from 15492 samples. We used only cell proteomics data sets that had been induced by a single drug. This totals 13738 samples. Proteomics data that were collected after a 24-hour incubation period was more useful according to previous studies.[11] For those drug-cell experiments that were measured at many separate time points, we kept only those that were analyzed after 24 hours. For samples that were only tested for less than 24 hours, we retained data that were incubated for the longest period of time. This resulted in a total of 8072 samples. We next filtered out control samples, cells, and drugs that were not relevant to the research, resulting in a total of 2268 samples. After averaging the signatures, the total sample size was 1341, consisting of 57 drugs and 73 cell lines. The initial antibody number was 549; we discarded those with a missing value of 100 percent and retained 512 proteins. Nonetheless, the percentage of missing values remained at about 60% as shown in Figure S2.

Perturbed transcriptomics data

The perturbed transcriptomics data were gathered from the Library of Integrated Network-based Cellular Signatures (LINCS) project.[40] This project collects an induced gene expression profile for 94 cell lines and more than 50,000 perturbations. The data we used were downloaded from https://github.com/njpipeorgan/L1000-bayesian. They are precomputed level 5 drug perturbed gene expression profiles generated with a more accurate and robust Bayesian-based peak deconvoluted approach.[43] As the 978 landmarks genes are determined to be more insightful in the drug perturbation study, we only included the gene expression profile of these 978 consensus signature genes. Furthermore, previous research had demonstrated that data quality influenced the predictive potential of the data, we only kept the most reliable data, as assessed by average Pearson correlation (APC) scores. The average Pearson correlation among biological replications was used to calculate the APC score. If the APC score for each

drug-cell-dosage-time combination is higher, it means the aggregated data is more trustworthy. We chose samples with a dose level of 10.0 μM, which produced 1427 (15 cell lines and 409 drugs) perturbations out of the 7121 total signatures.

Basal transcriptomics data

The basal transcriptomics dataset combines gene expression profiles of cell lines from CCLE,[32] 1305 cell line gene expression data obtained as part of the CCLE project were downloaded from DepMap https://depmap.org/portal/download (DepMap Public 20Q3).

Adverse drug reaction prediction data

We applied two adverse drug response datasets for the adverse drug reaction prediction task. The on-label adverse drug responses side effect resource (SIDER) dataset has 834 marketed drugs, 3,166 adverse drug response preferred terms, and 88,635 drug-ADR associations.[31] The off-label ADRs PharmGKB Offsides dataset from FDA adverse event report system (FAERS) has 684 drugs, 9,405 ADR terms, and 26,0238 drug-ADR associations.[30] For the ADR terms, we also removed the ADR terms that had only less than 10 drugs. The ADR terms used in SIDER and FAERS were labeled with the preferred terms from MedDRA v16.0.[44] For a fair comparison between different input features, a subset of drugs was selected for the model training and testing, as shown in Table S2.

Anti-cancer drug sensitivity prediction data

CCLE[32] contains the basal gene expressions of 1305 cancer cell lines using RNA-seq. We selected the cell lines with drug sensitivity information available from GDSC phase 1 and phase 2,[33,34] as well as the 978 landmark genes from LINCS project,[40] resulting in 680 cell lines and 370 chemicals in total for the following experiments. GDSC data can be downloaded from https://www.cancerrxgene.org/downloads/bulk_download, the version we used in this study was v8.2. The measurement in the experiments was the Z score which represents the Z score of the $LN(IC_{50})$ (x) comparing it to the mean (μ) and standard deviation (θ^2) of the $LN(IC_{50})$ values for the drug in question over all cell lines treated.

Baseline models

Neural fingerprints

Instead of applying Graph Isomorphism Networks (GINs) as the chemical embedding network as in TransPro, we built a baseline model with the original GCN for getting chemical neural fingerprints as the chemical features as demonstrated in this work.[22] The GCN takes a graph structure of a chemical compound as input and uses convolutional operations to update vector representations for each node (atom) in the graph (chemical compound). The chemical fingerprint is composed of the sum of the vectors of each node and then passed to the drug-specific DNN in Figure 1A. Except for the absence of the attention module and perturbed transcriptomics training, the rest of the model is essentially equal to Trans-Pro.

k-NN

Similar to TransPro, the k-NN model's input is a numerical representation of the chemical and cell line, but instead of data-driven representations for compounds using GNN, we use predetermined chemical fingerprints from PubChem[45] that are encoded as binary (bit) vectors that represent the presence or absence of particular substructures in chemicals, they are calculated with the Chemistry Development Kit.[46] In our settings, we experimented with deepChem fingerprints which have lengths (i.e. number of substructures) of 881. A perturbed proteomics profile is derived for an OOD chemical compound/cell line by averaging the proteomics perturbations of its nearest neighbors in the training set in the same setting. We experimented with varying the number of neighborhoods in the training set from five to fifteen, as well as with other measures of similarity, including cosine, correlation, Jaccard, and euclidean distance to report the best performance. The model was implemented using the Python Scikit-Learn library.

Random forest regressor

Random forest is an additional baseline that we used in our studies. It is a reliable bagging method (ensemble) that may be used to solve problems in both regression and classification. As with k-NN, it accepts predetermined fingerprints as chemical representations and cell lines, trains a large number of decision trees, weights the input features, and outputs the average of the individual trees' predictions of proteomics perturbation for a new chemical compound/cell line. We also experimented with different tree numbers during training in order to report the final findings with proper variance. Python Scikit-Learn was also used to build the model.

Vanilla neural network

Vanilla neural network is a simple two-layer fully connected neural network with a ReLu activation function, and it receives the same input as the baseline models discussed in k-NN and random forest. We explored a variety of hyper-parameter searches in order to obtain optimal performance. Likewise, we utilized the Scikit-Learn to implement this method.

TransPro modules

Graph Neural Network

The cell transcriptomics profile is compressed to a low-dimensional vector the encoder module. To build a drug-specific diff vector from the drug, the diff vector generation layer $E_drug - diff$ is employed. The drug-specific diff vector, we presume, represents the vector difference between the basal cell transcriptomics hidden vector and the perturbed transcriptomics hidden vector due to drug-induing. The difference vector in transcriptomics latent space is predicted using the drug graph neural network and the following feed-forward network. We are primarily interested in Graph Isomorphism Networks (GINs),[18] thus, the architecture of the backbone GNN Figure 1A is a five-layer GIN with 512 and 256 hidden units for MLPs in each layer,[20] more implementation details can be found in Table S1.

$$h_v^k = ReLU\left(MLP^k\left(\sum_{u \in N(v)} h_u^{k-1}\right) + \sum_{e=(v,u):u \in N(v)} h_e^{k-1}\right) \quad \text{(Equation 1)}$$

where N (v) is a set of nodes adjacent to v, and e = (v, v) represents the self-loop edge. Note that we eliminated the ReLU from the above equation or the last layer, i.e., k = K, so that h_v^k can take negative values. This is critical for pre-training methods based on the dot product, such as Context Prediction and Edge Prediction, because the dot product of two vectors would otherwise always be positive. The graph-level representation hG is computed by averaging the last layer's node embeddings, i.e.

$$h_G = MEAN\left(h_v^K \middle| v \in G\right) \quad \text{(Equation 2)}$$

We utilized RDKit to extract information about the drug's atoms and bonds from its SMILES string [16]. Then, using a feedforward neural network, the output h_G was used to construct the drug-specific diff vector.

$$Z_{diff} = f_{DNN}(h_G) \quad \text{(Equation 3)}$$

Cell line encoder and domain-specific decoders

We experimented with an encoder and decoder that were both fully connected to two-layer MLPs with one hidden layer followed by dropout. The activation function is ReLU, we allocate E to the encoder, and D_{trans}, D_{prot} to the decoders for transcriptomics perturbation and proteomics perturbation, respectively. It is worth noting that we did not add an activation function to the last layer because the prediction output was continuous.

Multi-head attention for drug-gene interaction network

We utilized a multi-head attention technique to quantify interactions between cell and chemical features, as illustrated in the Figure S4. In principle, each element of a set may be represented as a collection of three vectors: query, key, and value. A module for individual attention performs the mapping of queries and sets of key-value pairs to an output matrix.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad \text{(Equation 4)}$$

where Q, K, and V are corresponding matrices (sets) of queries, keys, and values, T is a transposition operation, and d_k is a scaling factor. Given a N-dimensional vector Q of chemical embedding and a M-dimensional vector K of cell line embedding, a NxM weight matrix will be trained to assign the importance of genes perturbed by the chemical. Another M-dimensional V of the cell line will be modified by the learned weight. By concatenating several individual attention modules, multi-head attention focuses on different representation sub-spaces. The drug-induced shifted hidden vector was calculated with the following equation:

$$Z_{drug-induced} = Z_{diff} + Attention(E(x_{basal-trans}), Z_{diff}) \quad \text{(Equation 5)}$$

where $x_{basal-trans}$ is the basal transcriptomics. The drug-induced hidden vector is then decompressed from transcriptomics latent space to transcriptomics profile with the D_{trans}:

$$y'_{pert-trans} = D_{trans}\left(Z_{drug-induced}\right) \quad \text{(Equation 6)}$$

Transmitter

The transmitter is a simple MLP to bridge the information between transcriptomics latent space and proteomics latent space. The drug-induced hidden vector is decompressed from transcriptomics latent space to proteomics profile with the *Transmitter* and D_{prot} :

$$y'_{pert-prot} = D_{prot}\left(Transmitter\left(Z_{drug-induced}\right)\right) \quad \text{(Equation 7)}$$

Adverse drug reactions prediction

The side effect prediction network is a three-layer feed-forward neural network with a ReLU activation function that took the cell line generated from the hidden space $Transmitter(Z_{drug-induced})$ as the input as follows:

$$Y_{side-effect} = W_2\left(ReLU\left(W_1 Transmitter\left(Z_{drug-induced}\right) + b_1\right)\right) + b_2 \quad \text{(Equation 8)}$$

Anti-cancer drug sensitivity prediction

We first trained the model with the proteomics/transcriptomics perturbation task. An additional DNN head with a three-layer feed-forward and a ReLU activation function was connected to either the pre-trained chemical-induced transcriptomics hidden state or chemical-induced proteomics hidden state in Figure 1. The final output variable was IC50 values of anti-cancer drug sensitivity. During the training stage, we kept all the previous modules unfrozen so the gradient descent would propagate all the way back to the beginning to achieve end-to-end training. We evaluated the performance with 3-fold cross-validation.

When the network was connected to the perturbed proteomics latent space $Transmitter(Z_{drug-induced})$, the object function was

$$Y_{IC50-w-pert-prot-emb} = W_2\left(ReLU\left(W_1 Transmitter\left(Z_{drug-induced}\right) + b_1\right)\right) + b_2 \quad \text{(Equation 9)}$$

When the network was connected to the perturbed transcriptomics latent space $Z_{drug-induced}$, the object function was

$$Y_{IC50-w-pert-trans-emb} = W_2\left(ReLU\left(W_1\left(Z_{drug-induced}\right) + b_1\right)\right) + b_2 \quad \text{(Equation 10)}$$

Model training

TransPro model training harbors two major training tasks, perturbed transcriptomics prediction and perturbed proteomics prediction. Both tasks share the GNN drug embedding network, the cell encoder, and the Multi-head attention interaction network. It was worth mentioning that the perturbed transcriptomics data was always retrained in the training dataset because we mainly focused on the task of perturbed proteomics prediction for the model evaluation. The detailed training procedure is in Algorithm 1.

Algorithm 1. TransPro procedure

Input: $\{x_{trans}^{(i)}\}_{i=1}^{n_{trans}}, \{x_{prot}^{(i)}\}_{i=1}^{n_{prot}}$
Require: $n_{training}$, number of training epochs
1: **for** *epoch* = 1 *to* $n_{training}$ **do**
2: **for** $\{x_{trans}\} in \{x_{trans}^{(i)}\}_{i=1}^{n_{trans}}$ **do**
3: Update $E, E_{drug_{diff}}, Attention, D_{trans}$ *with* $L_{transpert}$
4: **end for**
5: **for** $\{x_{prot}\}$ *in* $\{x_{prot}^{(i)}\}_{i=1}^{n_{prot}}$ **do**
6: Update $E, E_{drug_{diff}}, Attention, transmitter, D_{prot}$ *with* $L_{protpert}$
7: **end for**

Loss function

When calculating the perturbed proteomics reconstruction loss, we masked out all the missing values in the labeled data and applied weighted mean square error as our main loss function between predictive proteomics and ground truth labels. Both n_{prot} and $y_{pert-prot}$ excluded the missing values in order to calculate the final loss. Similarly, we excluded missing data from our evaluation procedure.

$$L_{prot} = \frac{1}{n_{prot}} \sum_{i=1}^{n_{prot}} \left\| \left(y_{pert-prot} - y'_{pert-prot}\right) \right\|_2^2 \quad \text{(Equation 11)}$$

The perturbed transcriptomics prediction loss function is a regular mean square error loss as follows:

$$L_{trans} = \frac{1}{n_{trans}} \sum_{i=1}^{n_{trans}} \left\| \left(y_{pert-trans} - y'_{pert-trans}\right) \right\|_2^2 \quad \text{(Equation 12)}$$

QUANTIFICATION AND STATISTICAL ANALYSIS

Performance evaluation

Throughout the experiments, we focused on Pearson correlation and Spearman correlation as evaluation metrics for expression profile prediction and anti-cancer drug sensitivity prediction. Correlation scores for assessing the relationship between ground truth and predicted gene expression data have been shown to be more effective than error measures in comparison to microarray data analysis. We used both correlation coefficients to evaluate the performance of TransPro due to the different information they may provide for downstream tasks. When comparing predicted and true gene expression values, Pearson correlation is preferable to Spearman's correlation. Spearman correlation is better suited for comparing gene expression rankings, which offers further practical information for drug sensitivity prediction such as the ranking of IC_{50}.

AUROC and AUPRC were the measurements we used to evaluate the multi-labeled binary classifier for the adverse drug reactions prediction, and we used the macro average for both. Although ROC-AUC is a widely used statistic to evaluate a classifier's performance, ROC curves give a more optimistic impression of the model, particularly when working with an imbalanced dataset.[47] Because only true positive rate and false-positive rate are taken into account in the ROC, using a ROC curve with an imbalanced dataset is misleading and may result in an incorrect evaluation of the model.[48] They are independent of the distribution of classes.[49]