Transformer-based deep learning integrates multi-omic data with cancer pathways

Authors:

Zhaoxiang Cai[1], Rebecca C. Poulos[1], Adel Aref[1], Phillip J. Robinson[1], Roger R. Reddel[1], Qing Zhong[1*]

Affiliations:

[1] ProCan®, Children's Medical Research Institute, Faculty of Medicine and Health, The University of Sydney, Westmead, NSW, Australia.

Notes:

* Correspondence.

Keywords:

Multi-omics; Deep Learning; Cancer; Model Explanation, Cancer Pathway

Abstract

Multi-omic data analysis incorporating machine learning has the potential to significantly improve cancer diagnosis and prognosis. Traditional machine learning methods are usually limited to omic measurements, omitting existing domain knowledge such as the biological networks that link molecular entities in various omic data types. We develop a Transformer-based explainable deep learning model, DeePathNet, which integrates cancer-specific pathway information into multi-omic data analysis. Using a variety of big datasets, including ProCan-DepMapSanger, CCLE and TCGA, we show that DeePathNet outperforms traditional methods for the prediction of drug response and classification of cancer type and subtype. Combining biomedical knowledge and state-of-the-art deep learning methods, DeePathNet enables biomarker discovery at the pathway level, maximising the power of data-driven approaches to cancer research. DeePathNet is available on GitHub at https://github.com/CMRI-ProCan/DeePathNet.

Introduction

Multi-omic analysis of diverse data types enables researchers to gain insights into tumour biology and to identify new and robust therapeutic targets[1]. One major goal of multi-omic analysis by machine learning is to predict the cancer treatment strategies that are best suited to individuals in the context of precision medicine. To date, advances in precision medicine applications in cancer have been driven primarily by the availability of relevant technologies to measure genomic and transcriptomic data[2]. Recent advances in mass spectrometry have enabled large-scale proteomic data to be acquired in human cancer tissues, organoids and cell lines, contributing to expanded possibilities for multi-omic data analysis[3–5].

Multi-omic analysis enables increased understanding of tumour biology and the identification of more robust therapeutic targets[5]. A variety of multi-omic studies have led to the improved detection of intra-tumour heterogeneity, identification of novel therapeutic targets, as well as more robust diagnostic and predictive markers[6–9]. Many of these discoveries would not have been possible by analysing any single omic data type alone. However, performing multi-omic analysis presents computational challenges due to the large number of data generated by high-throughput instruments and the limitations of existing multi-omic data integrative methods[10,11].

To address this, a plethora of machine learning methods have been developed for integrating large-scale multi-omic data [6–8,11–15]. For example, moCluster[12] integrates multi-omic data based on joint latent variable models, showing performance superior to previous methods such as iCluster[13] and iCluster Bayes[14]. Likewise, mixOmics[15] provides various options for multi-omic data integration, aiming to find common information between different omic data types. These models solely take omic measurements as the input and do not consider existing biomedical knowledge that links different omic data types together, such as the regulatory networks. Regulatory networks exist in cells to control the expression

levels of different gene products, through collections of functionally interacting protein or RNA macromolecules[16]. However, models that incorporate existing biomedical knowledge in addition to computational inference have the potential to better capture the interactions that drive biomarker associations, and to increase the predictive power and modelling capacity of these algorithms[11].

Several studies have attempted to incorporate existing biomedical knowledge into multi-omic models using deep learning[17]. DCell[18] and DrugCell[19] combine the neural network architecture with known gene ontology information, but they only support the use of gene deletions or mutation as the input. EMOGI[20] was designed based on graph neural networks[21] and integrates protein-protein interaction (PPI) networks with multi-omic data to predict cancer genes, but its network architecture cannot be easily generalised to other tasks. Besides, gene ontology information and PPI networks used in these models do not precisely reflect cancer-specific information. Therefore, integrating cancer pathways[22] into multi-omic data analysis by deep learning for general tasks, such as drug response prediction and cancer type or subtype classification, remains an open research topic.

To address this gap, we developed DeePathNet, a Transformer-based[23] explainable deep learning method that inputs multi-omic data alongside knowledge of cancer pathways. The Transformer is used primarily in the fields of natural language processing and computer vision, by adopting the mechanism of self-attention[23]. In molecular biology, the Transformer-based model AlphaFold has successfully predicted protein structures based on amino acid sequences[24]. However, the Transformer has not yet been incorporated into multi-omic cancer data analysis. Here, we apply DeePathNet by using a Transformer module to integrate several large-scale multi-omic datasets with cancer pathways, allowing more complex patterns to be learned. By comprehensively evaluating multiple datasets with three prediction tasks and a

range of metrics, we demonstrate that the predictive power of DeePathNet is superior to that of traditional machine learning methods.

Results

Overview of DeePathNet

DeePathNet was developed to model biological pathways using a Transformer-based deep learning architecture with both multi-omic data and cancer pathway information as the input **(Fig. 1a)**. The performance of DeePathNet was evaluated on drug response prediction, and cancer type and subtype classification.

DeePathNet consists of three major steps. It starts with a pathway encoder to summarise features from an arbitrary number of omic data types into cancer pathways (Step 1; **Fig. 1b**), and then uses a Transformer encoder to model the interactions between these pathways (Step 2). This is followed by a multi-layer perceptron (MLP) that can be adapted to different prediction tasks (Step 3).

In Step 1, the neural network architecture is constructed based on the LCPathways dataset[22], which contains 241 literature-curated pathways encompassing 3,164 cancer genes. The pathway encoder then uses a fully connected layer to project the multi-omic data (Omics 1–*m*) from genes (Gene 1–*n*) onto a 512-dimension pathway vector that represents one of the cancer pathways (Pathway 1–*p*; **Supplementary Fig. 1a**, see **Methods**). With this architecture, the pathway encoder allows DeePathNet to capture interactions across different omic data types.

In Step 2, an enhanced version of the Transformer module is developed to encode the interactions between cancer pathways (**Supplementary Fig. 1b,** see **Methods**). First, a dropout layer is used to train only half of the pathways at each iteration, which prevents the model from focusing on specific pathways that may not generalise well to a test dataset. Then, two blocks of the original Transformer module[23] are used, which contains a list of recurring layers with each layer comprising a sequence of layer normalisation, multi-head

self-attention, and a MLP. The Transformer also enables dynamic modelling of the complex relationship between cancer pathways, thus avoiding the generation of fixed weights for the different input, as is the case in traditional machine learning.

In Step 3, a MLP is used to map the encoded pathway vectors to output neurons, which allows the knowledge learned by the Transformer module to be adapted to general prediction tasks.

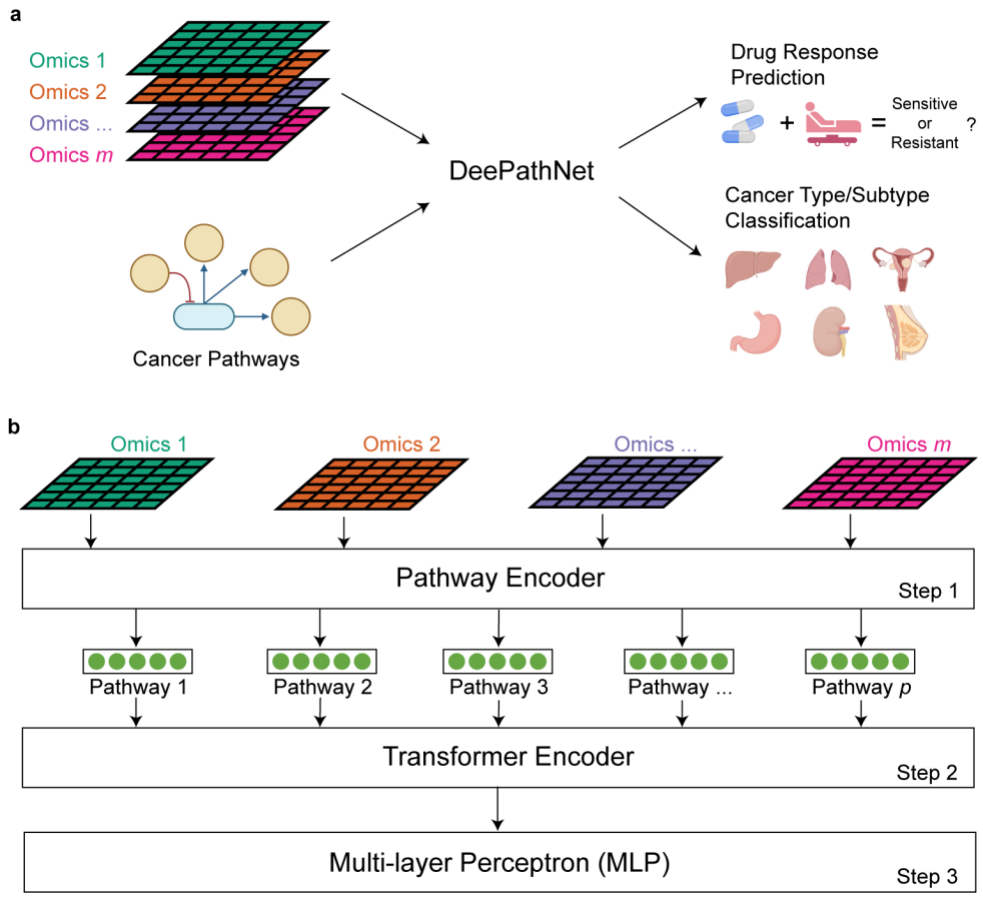


Fig. 1 | Overview of DeePathNet. a, DeePathNet has its network architecture built using the LCPathway dataset and takes multi-omic data as the input to model pathway interactions and predict drug responses or classify cancer types and subtypes. **b,** DeePathNet architecture supports any number of omic data types as the input. Step 1: DeePathNet encodes multi-omic information into an arbitrary number of cancer pathways. Step 2: DeePathNet uses a Transformer encoder to learn the interactions between these pathways. Step 3: The encoded pathway vector is passed into a MLP for the prediction. Circles represent neurons in a neural network. Arrows represent the direction of information flow.

DeePathNet predicts drug response

We first assessed the predictive performance of DeePathNet on a regression task by benchmarking it against random forest[25], elastic net[26], principal component analysis (PCA), mixOmics[8] and moCluster[12] to predict the responses of anti-cancer drugs to cancer cell lines. These six methods were evaluated using data from the Cell Lines Project (CLP)[27] and the Cancer Cell Line Encyclopedia (CCLE)[28], the two largest publicly available multi-omic cancer cell line datasets (**Supplementary Table 1,** see **Methods**). Gene mutation, copy number variation (CNV) and gene expression data from the two datasets were used as the input. For drug response data, we retrieved the half-maximal inhibitory concentration (IC_{50}) from the Genomics of Drug Sensitivity in Cancer (GDSC)[27] database. For each method, six experimental setups were assessed, comprising two datasets and three evaluation metrics, namely coefficient of determination (R^2), mean absolute error (MAE) and Pearson correlation coefficient (Pearson's *r*) between predicted and actual IC_{50} values.

In DeePathNet, 241 pathway encoders were constructed (**Supplementary Fig. 1a**) to summarise the omic data into pathway vectors defined by the LCPathways[22]. These vectors were then fed into the Transformer module to model the interactions between cancer pathways (**Supplementary Fig. 1b**). Default hyperparameters were used for all six methods (see **Methods**). Omic data were combined using early integration[11] for random forest and elastic net. Middle integration[11] was used for PCA, moCluster and mixOmics. PCA and moCluster were coupled with random forest for predictions[11] (see **Methods**).

To quantitatively and reliably compare the six methods, five-fold cross-validation was repeated five times at random, yielding 25 error measures for each of the R^2, MAE and Pearson's *r* metrics. The mean and 95% confidence interval (CI) of the evaluation metrics

was reported, serving as an estimate of the generalisation error. We observed that DeePathNet had significant and consistently better performance in drug response prediction than the other five methods that do not incorporate cancer pathway information (**Fig. 2a-f**, *p*-value < 0.0001, two-tail paired Student's *t*-test, **Supplementary Table 2**). By ranking the methods according to the mean measures for each setup, we found that random forest was the second-best performing method(**Fig. 2g**). To investigate whether drug responses that had relatively lower predictive accuracy by DeePathNet were also challenging for other methods, the correlations between DeePathNet and the other five methods were calculated. We found that the predictive performance of the paired methods was highly concordant (Pearson's r > 0.9), with DeePathNet consistently outperforming the other five methods (**Supplementary Fig. 2**).

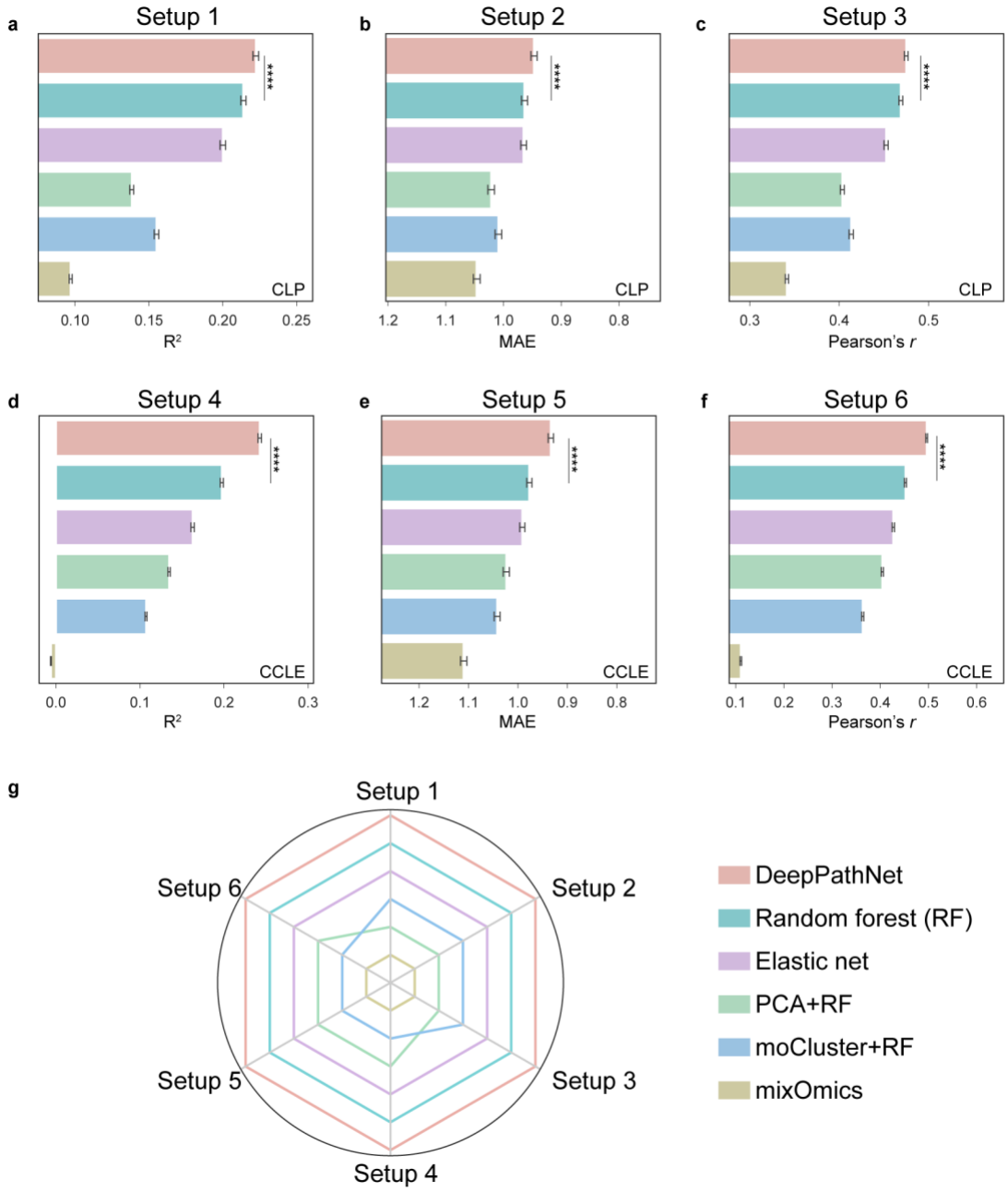


Fig. 2 | Performance evaluation of drug response prediction by cross-validation. a-f, Bar plots showing predictive performances across six experimental setups on the CLP and CCLE datasets by three evaluation metrics: R^2, MAE (inverted on the horizontal axis), and Pearson's *r*. A higher value represents better performance. Error bars are derived from cross-validation, representing 95% confidence intervals of the mean. **** indicates *p*-value < 0.0001 by two-tail paired Student's *t*-test, only showing significance between the first- and second-best performing methods. **g,** Radar plot showing the ranks of each model across the six experimental setups. A larger enclosed area represents better performance. **a-g**, The six methods are colour coded as in **g**.

To evaluate the generalisation error using an independent test set, we trained DeePathNet on the CLP dataset (**Supplementary Table 1**) and tested the final model by predicting drug responses in the CCLE dataset (**Supplementary Table 1**). Cancer pathway information was integrated in the same way as described above and random forest was trained as a baseline model. The test performance for all 549 GDSC anti-cancer drugs was summarised for both DeePathNet and random forest. DeePathNet achieved a statistically significant higher predictive performance than random forest across all three metrics (**Fig. 3a**, *p*-value < 0.0001, two-tail paired Student's *t*-test, **Supplementary Table 3**).

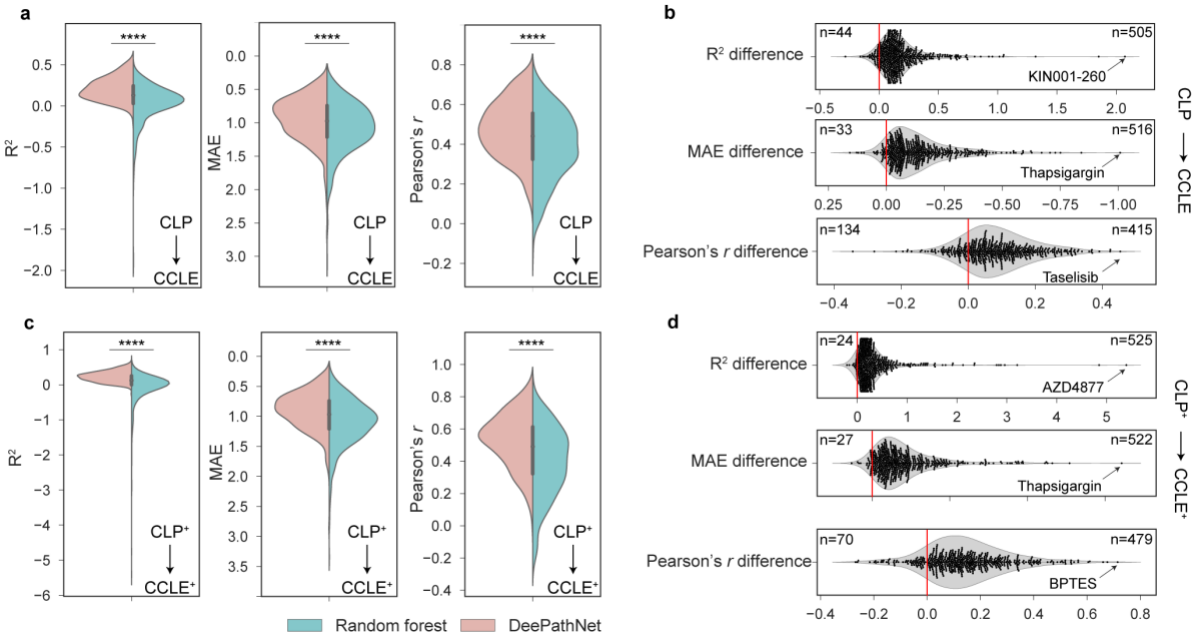


Fig. 3 | Generalisation error of DeePathNet and random forest for drug response prediction. a, Violin plots showing predictive performances of DeePathNet and random forest using CLP as the training set and evaluated on the independent CCLE test set across the 549 GDSC drugs.The vertical axis is inverted for MAE. **** indicates *p*-value < 0.0001 by two-tail paired Student's *t*-test. **b,** Violin and swarm plots showing the performance difference in R^2 (upper), MAE (middle) and Pearson's *r* (lower) between DeePathNet and random forest for each drug. A drug is more accurately predicted by DeePathNet when it exhibits a positive value for the R^2 or Pearson's *r* difference, or a negative value of the MAE difference (the horizontal axis is inverted for MAE). The numbers of drugs that are more accurately predicted by DeePathNet or random forest are annotated on the upper right and left of the plot, respectively. The name of the drug that achieved the largest improvement with DeePathNet is annotated for each metric. **c** and **d**, Similar to **a** and **b**, but using CLP^+ as the training set and $CCLE^+$ as the independent test set.

To compare the predictive performance of DeePathNet with random forest for each drug, the difference of R^2 between DeePathNet and random forest was measured. Here, 92% (505/549) of drugs had positive values, indicating superior predictive performance from DeePathNet over random forest. Similarly, 94% (516/549) drugs and 76% (415/549) of drugs exhibited improved results by MAE and Pearson's *r*, respectively **(Fig. 3b).** This demonstrates that DeePathNet consistently achieved better predictive performance than random forest for most anti-cancer drugs. The drug that obtained the largest R^2 improvement by DeePathNet was KIN001-260 (**Fig. 3b**), which was poorly predicted by random forest and caused the long tail in the distribution of values (**Fig. 3a**). Drugs that had the largest improvement in MAE and Pearson's *r* with DeePathNet were thapsigargin and taselisib (**Fig. 3b**).

Next, we extended our analysis by including two proteomic cell line datasets from ProCan-DepMapSanger[3] and CCLE[29]. ProCan-DepMapSanger is a recently published pan-cancer proteomic dataset of 949 human cell lines generated by our team, supplementing the CLP with proteomic information. DeePathNet and random forest were trained on the combined CLP and ProCan-DepMapSanger datasets (CLP^+; **Supplementary Table 1**), with the final model tested on the expanded CCLE dataset that includes additional proteomic measurements ($CCLE^+$; **Supplementary Table 1**). Pathway information was integrated in DeePathNet as described above. DeePathNet yielded significantly higher test performance than random forest across all three metrics when predicting the 549 GDSC anti-cancer drugs (**Fig. 3c**, **Supplementary Table 3**). Analysing the predictive performance for each drug, DeePathNet also provided significant improvement for the majority of anti-cancer drugs compared with random forest (**Fig. 3d**). The drugs that had the largest improvement by DeePathNet were AZD4877, thapsigargin and BPTES, measured by the differences of R^2, MAE and Pearson's *r*, respectively (**Fig. 3d**).

To investigate which types of drugs were most accurately predicted by DeePathNet, we grouped the 549 drugs by their canonical target cellular pathways. Drugs targeting ABL signalling and ERK MAPK signalling pathway had the highest mean Pearson's *r* between predicted and actual IC_{50} values (**Supplementary Fig. 3a**). The top 20 most accurately predicted drugs and their pathways are reported in **Supplementary Fig. 3b**.

Taking these observations together, we demonstrated DeePathNet increased predictive performance through several benchmarking analyses in predicting responses to several drugs targeting various signalling pathways.

DeePathNet classifies cancer types

To evaluate DeePathNet on a classification task, we used publicly available data from The Cancer Genome Atlas (TCGA)[30] to classify primary cancer types. Gene mutation, CNV and gene expression features were used as the omic data input to train DeePathNet models to classify each of the 6,356 samples into one of 23 cancer types (see **Methods**). A total of seven metrics were used across the analysis to ensure reliable evaluation. The metrics are accuracy, macro-average F_1-score, precision, recall (sensitivity), area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC) and stability (see **Methods**). LCPathways was integrated in the same way as described for the drug response prediction. For benchmarking, elastic net was replaced with *k*-nearest neighbours (*k*-NN)[31] because elastic net does not support classification. For all six methods, feature integration and hyperparameter settings were identical to the drug response prediction.

In the absence of an independent dataset comprising the 23 cancer types, cross-validation was performed for the six methods on the TCGA dataset, and the mean and 95% CI of the evaluation metrics were reported as an estimate of the generalisation error. DeePathNet consistently outperformed the other five machine learning methods by accuracy, macro-average F_1-score (**Fig. 4a**, **Supplementary Table 4**). In contrast, other methods such as mixOmics only performed well in one metric, indicating that these methods may be suitable for certain scenarios but can not generalise well across datasets (**Fig. 4a**). Assessing the performance of each method using a set of four metrics including accuracy, macro-average F_1-score, AUROC and stability, showed that DeePathNet was consistently top ranked, followed by random forest (**Fig. 4b, Supplementary Table 4**).

To further investigate DeePathNet's performance for each cancer type, the predicted and actual cancer type for each sample was visualised using a confusion matrix, with the number of samples, precision and recall annotated (**Fig. 4c**). DeePathNet achieved a recall of over 0.95 for most cancer types, with acute myeloid leukemia (LAML), pancreatic adenocarcinoma (PRAD), and thyroid carcinoma (THCA) as the top three most accurately classified cancer types. Rectum adenocarcinoma (READ) was the cancer type with the lowest recall, having 46% of the samples incorrectly classified as colon adenocarcinoma (COAD). The latter outcome is unsurprising because the colon and rectum are adjacent tissue types that share highly similar features, with these two cancer types often grouped together[32] and are treated with similar chemotherapeutic regimens[32]. The cancer type exhibiting the second-lowest recall was stomach adenocarcinomas (STAD), with 19% of STAD samples incorrectly classified as esophageal carcinoma (ESCA). This can be explained by their similar histopathology and the anatomical proximity of STAD and ESCA[33]. Next, AUROC and AUPRC were examined for each cancer type, both displaying high performances for all cancer types, with the exception of AUPRC for READ, due to the tissue proximity of READ to COAD. (**Supplementary Fig. 4a** and **Supplementary Fig. 4b**).

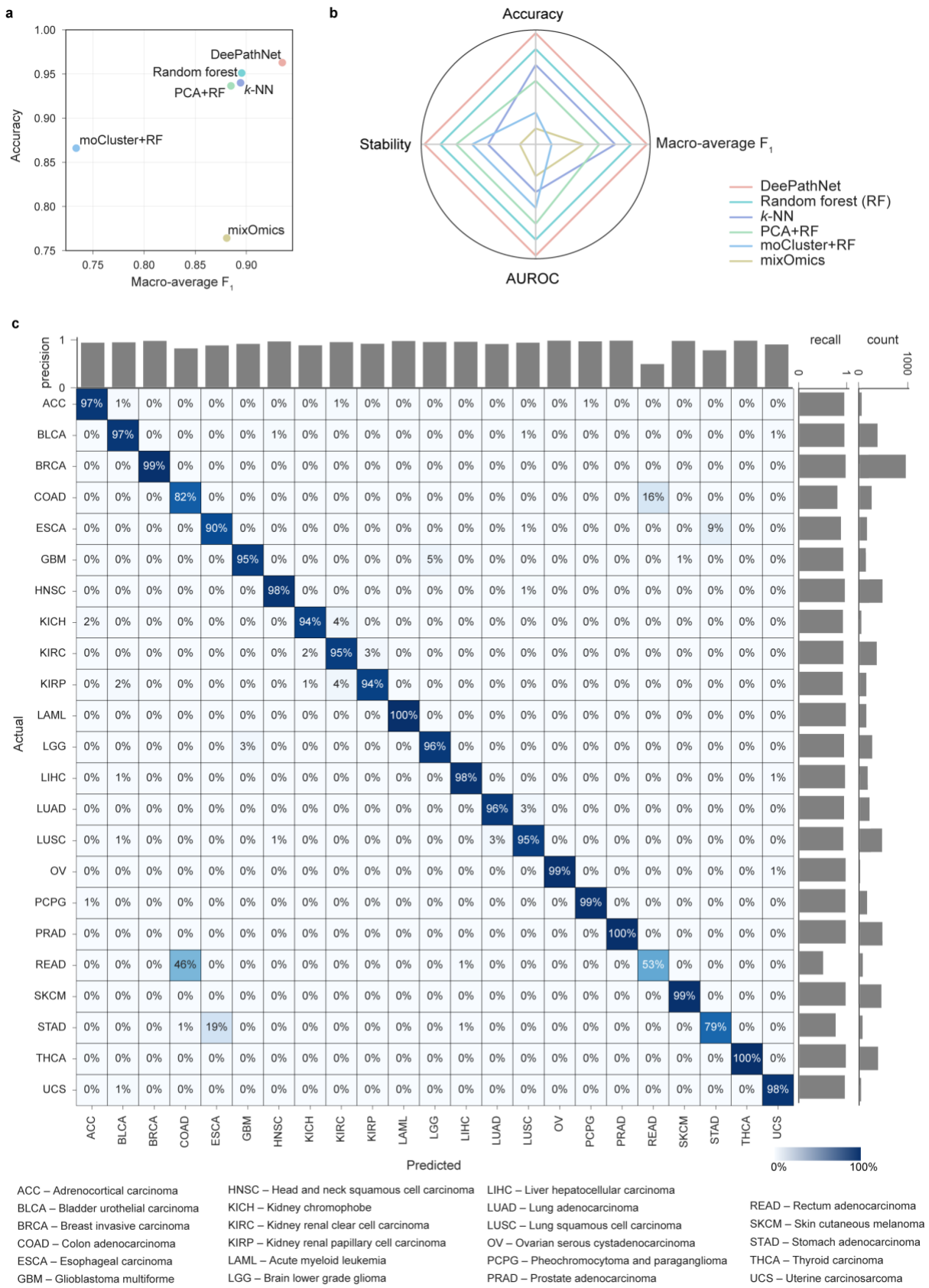


Fig. 4 | Performance evaluation of cancer type classification. a, Model comparison using cross-validation on the TCGA dataset. The x-axis represents macro-average F_1-score, and the y-axis denotes accuracy. **b,** Radar chart showing the model ranks across the set of four metrics. A larger enclosed area

indicates better predictive performance. **c,** Confusion matrix for the classification of 23 cancer types. Columns denote predicted labels, and rows represent actual labels. The percentage shown represents the proportion of predictions made for the corresponding cancer type, with each row summing to 1. The diagonal represents correct predictions for each cancer type, with the percentage indicating the recall. Bar plots show precision (horizontal axis), recall (vertical axis, leftmost) and number of samples (vertical axis, rightmost) per cancer type.

DeePathNet classifies breast cancer subtypes

Gene mutation, CNV and gene expression features were used to train DeePathNet models for the classification of five breast cancer subtypes (Luminal A, Luminal B, HER2+, Basal, Normal-like) according to the Prediction Analysis of Microarray 50 (PAM50)[34]. A total of 974 breast cancer samples from the TCGA dataset were used for training, and a breast cancer cohort of 122 samples from Clinical Proteomic Tumor Analysis Consortium (CPTAC) was included as an independent dataset to evaluate the generalisation error.

Cross-validation for all six methods was first performed on the TCGA dataset, reporting the mean and 95% CI of the evaluation metrics as an estimate of the generalisation error. DeePathNet provided a substantial improvement over the other methods in terms of accuracy and macro average F_1-score (**Fig. 5a**, **Supplementary Table 5**). The performance gain in AUROC was relatively minor but statistically significant (Student's *t*-test *p*-value < 5×10^{-4}) (**Supplementary Table 5**). The methods were then ranked according to the same set of four metrics as in cancer type classification. DeePathNet achieved the best performance in all four metrics, with random forest ranked as the second best overall (**Fig. 5b**). Other methods showed inconsistent performance rankings across different metrics, demonstrating the necessity of using multiple evaluation metrics for a comprehensive evaluation.

To evaluate the generalisation error, a DeePathNet model was trained on the TCGA breast cancer cohort, with the final model tested on the independent CPTAC breast cancer cohort. Benchmarked against random forest, DeePathNet yielded a much lower generalisation error on the independent test set (**Fig. 5c**, **Supplementary Table 6)**. Next, the generalisation error of DeePathNet was assessed for each subtype by a confusion matrix. DeePathNet achieved the highest precision and recall in classifying the Basal subtype (96.6%, **Fig. 5d**), with most tumours in this subtype being high-grade with a poor prognosis[35]. The most difficult subtype

to classify was Normal-like, where three out of the five Normal-like samples were incorrectly classified as Luminal A (**Fig. 5d**). Luminal A and Normal-like subtypes are traditionally difficult to distinguish as they share the same immunohistochemistry markers[35]. The Normal-like subtype is less frequently used in clinics[36]. Further analyses by AUROC (**Supplementary Fig. 5a**) and AUPRC (**Supplementary Fig. 5b**) demonstrated DeePathNet's high predictive performance for each subtype.

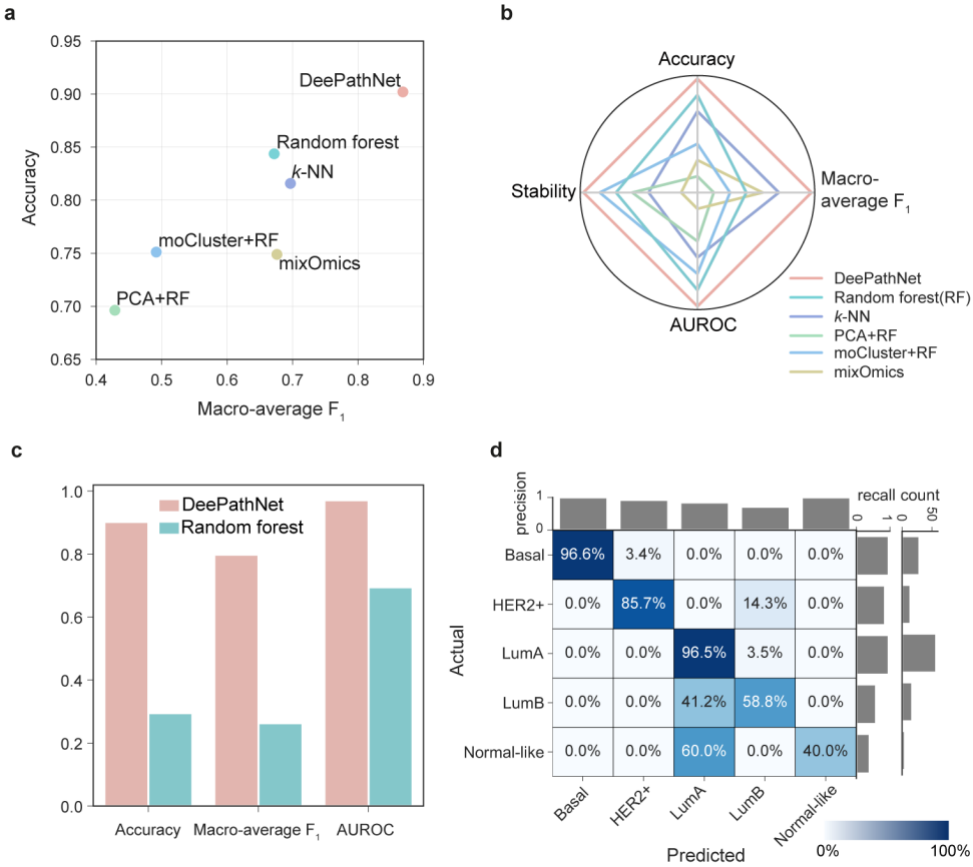


Fig. 5 | Performance evaluation of breast cancer subtype classification. a, Model evaluation by cross-validation. The x-axis represents macro-average F_1-score, and the y-axis represents accuracy. **b,** Radar chart showing the model ranks across the set of four metrics. A larger enclosed area represents better classification performance. **c,** Performance metrics showing generalisation errors for DeePathNet and random forest when using CPTAC data as the independent test set. **d,** Confusion matrix showing generalisation errors when using CPTAC data as the independent test set. Statistics are annotated in the same way as described in **Fig. 4c**.

DeePathNet provides model explanation

The DeePathNet model is explainable at both omic and pathway levels by using feature importance derived from SHapley Additive exPlanations (SHAP)[37] and Layer-wise Relevance Propagation (LRP)[38]. SHAP attributes the prediction to all features and assigns each feature an importance value, while LRP assumes that the classifier can be decomposed into several layers of computation, with these layers being parts of the feature extraction. Thus, both SHAP and LRP are post-hoc model explanation approaches that establish relationships between feature values and the predictions after DeePathNet is trained. Breast cancer subtype classification was used to demonstrate model explanation.

To explain the model at the omic level, SHAP was used to calculate feature importance. Specifically, feature importance was computed and visualised for the top five genes as stack bar plots comprising each omic data type for each breast subtype (**Fig. 6a**). DeePathNet was able to identify known biomarker genes as top features, such as ESR1, ERBB2 and KRT17, whose gene expression is routinely used to determine the PAM50 subtypes in the clinic[34] (**Fig. 6a**). Most genes had their high feature importance attributable to transcriptomic data (**Fig. 6a**), consistent with the fact that PAM50 classifications are RNA-based subtypes[34].

To explain the models at the pathway level, LRP was used to calculate feature importance. Since the cancer pathways are represented as an encoded vector that summarises multi-omic information, feature importance of a cancer pathway is computed for all omic data types jointly. For each cancer subtype, the top five pathways with the highest feature importance values were ranked (**Fig. 6b**). DeePathNet identified the FOXM1 transcription factor network as the most important pathway for predicting all PAM50 subtypes (**Fig. 6b**). FOXM1 shows distinct patterns of expression in different breast cancer subtypes and is seen as a promising candidate target in breast cancer treatment[39]. FOXM1 is also an adverse

prognostic factor of survival in Luminal A and B subtypes[40]. The ARF6 pathway was shown to be overexpressed in triple negative breast cancer and to be associated with breast cancer invasion and metastasis[41]. Similarly, Notch Signalling pathways are involved in cell proliferation, apoptosis, hypoxia and epithelial to mesenchymal transition and were found to be over-expressed in HER2+ positive and triple-negative breast cancer[42].

Taken together, these findings suggest that DeePathNet provides reliable model explanation with a strong biological basis by providing feature importance at both the omic and pathway level.

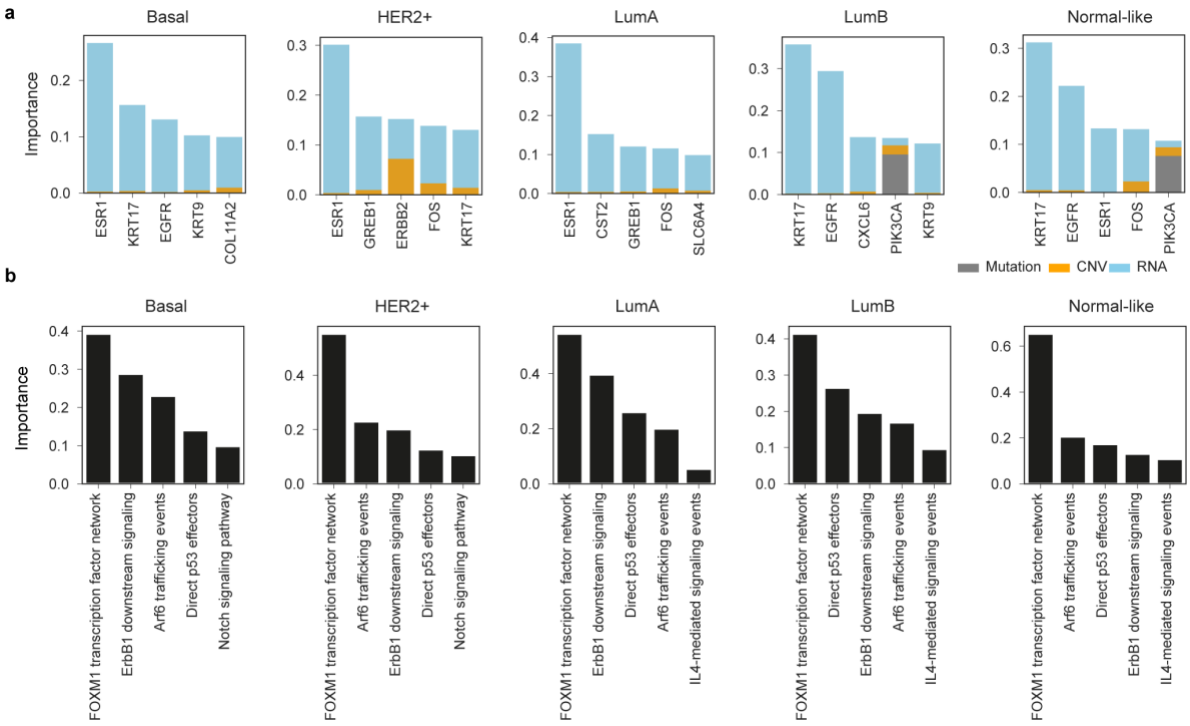


Fig. 6 | DeePathNet model explanation by omic level and pathway level feature importance. a, Stacked bar plots showing the omic level feature importance of the top five genes for each omic data type (indicated by grey, yellow and blue colour). **b,** Bar plots showing the DeePathNet pathway level feature importance of the top five pathways.

Discussion

DeePathNet is a Transformer-based deep learning model that overcomes the limitation of existing machine learning approaches that do not consider known cancer biology. DeePathNet integrates multi-omic data with cancer pathway knowledge to accurately predict drug responses and classify cancer types and subtypes. The self-attention mechanism of the Transformer module dynamically models the interdependency between pathways, thus capturing regulatory effects across different biological processes and the effects of dysregulation.

The predictive performance of DeePathNet was evaluated by one regression and two classification tasks. The evaluation was conducted on a larger scale than previous similar studies[12,15], using multiple big datasets and a range of metrics with both cross-validation and independent testing. Incorporating cancer pathway information, DeePathNet outperformed other machine learning methods that only use omic data as input features. A low generalisation error when validating DeePathNet models on independent datasets suggests that DeePathNet work well even when different experimental protocols were implemented between these independent datasets. DeePathNet provides model explanations at the pathway level, which has not yet been accomplished by other multi-omic integration tools for the prediction of drug response and classification of cancer type and subtype. DeePathNet was able to highlight known biomarkers when predicting breast cancer subtypes, including ESR1, ERBB2 and the FOXM1 network pathways. This suggests that other top-ranked genes and pathways may provide novel insights into cancer biology and drug discovery.

Despite these comprehensive evaluations, this study only concentrated on a limited number of omic data types because large-scale studies of some omic data types are still in their infancy[29,43,44]. As large proteomic and metabolomic datasets become increasingly

available, the predictive power of DeePathNet can improve further, because deep learning is likely to obtain performance boost with increased amount of data[45].

In conclusion, DeePathNet combines multi-omics, deep learning and existing biological knowledge to predict cancer phenotypes accurately with a model explanation. The application of DeePathNet may lead to more accurate diagnosis and prognosis, and will facilitate researchers to understand unknown cancer mechanisms and prioritise putative drug targets.

Methods

Multi-omic and drug response data collection. For drug response prediction, multi-omic data were retrieved from 941 CLP[27] and 696 CCLE cell lines[28]. In total, 19,099 gene mutation, 19,116 CNV and 15,320 gene expression features are in the CLP, and 18,103 gene mutation, 27,562 CNV and 19,177 gene expression features are in the CCLE.

For drug response prediction analysis with proteomic data, the ProCan-DepMapSanger dataset[3] was added to the CLP (CLP + ProCan-DepMapSanger = CLP^+) and the CCLE's proteomic dataset[29] was also used (CCLE + CCLE proteomic data = $CCLE^+$). The ProCan-DepMapSanger and CCLE proteomic datasets contain 8,498 and 12,755 protein features, respectively. The combined datasets have 910 and 292 cell lines for CLP^+ and $CCLE^+$, respectively. No additional processing was performed on the datasets (**Supplementary Table 1)**.

For cancer type and subtype classification, multi-omic data from TCGA cohorts were retrieved using TCGA-assembler 2[46]. In total, 6,356 samples were collected, containing 31,949 features from gene mutation, 23,529 features from CNV and 20,435 features from gene expression. In addition, multi-omic data from 122 breast cancer samples were retrieved from a CPTAC breast cancer cohort[47], containing 11,877 features from gene mutation, 23,692 features from CNV and 23,121 features from gene expression. For breast cancer subtype classification, the PAM50 classification (Luminal A, Luminal B, HER2+, Basal and Normal-like) was retrieved from the TCGA and CPTAC datasets (**Supplementary Table 1)**.

Overview of DeePathNet. DeePathNet has a pathway encoder (Step 1), a Transformer encoder (Step 2) and a MLP (Step 3).

In Step 1, DeePathNet encodes multi-omic information into cancer pathways, defined by the 241 cancer pathways in LCpathways[22]. Let $g_{mutation} \in \{0,1\}$ represent the mutation, $g_{CNV} \in \mathbb{R}$ the CNV, $g_{RNA} \in \mathbb{R}$ the gene expression, and $g_{prot} \in \mathbb{R}$ the protein intensity of a gene g. Then the vector that contains omic features for a pathway that contains n genes with four omic data types, is defined as:

$$\boldsymbol{a}_{omics} = \left[g^1_{mutation}, g^1_{CNV}, g^1_{RNA}, g^1_{prot}, \dots, g^n_{mutation}, g^n_{CNV}, g^n_{RNA}, g^n_{prot}\right]$$

Next, the vector $\boldsymbol{a}_{omics}$ is encoded into the pathway vector $\boldsymbol{a}_{encoded}$ via a MLP. Here, the notation is converted into the matrix form to include the number of samples. Thus, for N samples, the total features from the four omic data types for a pathway can be represented as a matrix A_{omics} of dimension $N \times 4n$. DeePathNet then uses a fully connected layer to encode these omic features into an encoded pathway matrix $A_{encoded}$, calculated as:

$$A_{encoded} = A_{omics}W^T + B$$

where W and B represent the learnable weights matrix and bias term in the fully connected layer. The dimension of the weight matrix W is set as $512 \times 4n$. The dimension of both bias B and $A_{encoded}$ is $N \times 512$. In total, 241 cancer pathways were used and 241 matrices $A^1_{encoded}, A^2_{encoded}, \dots, A^{241}_{encoded}$ are combined as a tensor $\boldsymbol{A}_{encoded}$ with a dimensionality of $N \times 512 \times 241$. $\boldsymbol{A}_{encoded}$ is used as the input into the Transformer encoder (**Supplementary Fig. 1**).

In Step 2, DeePathNet uses a Transformer encoder to learn the interdependence between regulatory pathways in cancer. In contrast to the general attention mechanism that models the interdependence between the input and target, self-attention is used by the Transformer module to model interdependence within the input[48] (i.e., features from the multi-omic data). The Transformer encoder starts with a dropout layer with a probability of 0.5 on the 241 cancer pathways, ensuring that on average half of the pathways are dropped out during training to prevent potential overfitting. The set of selected pathways is sampled independently for each training batch, allowing different pathways to be used. The Transformer block was configured the same way as the original version[23], denoted as $Transformer$ below. Since the Transformer encoder contains recurrent layers, we use a superscript with parenthesis to represent the $\boldsymbol{A}_{encoded}$ at different layers, where $\boldsymbol{A}^{(0)}_{encoded}$ represents the data before entering the first layer. After the first layer of the Transformer block, $\boldsymbol{A}^{(0)}_{encoded}$ becomes $\boldsymbol{A}^{(1)}_{encoded}$ as follows:

$$\boldsymbol{A}^{(1)}_{encoded} = Transformer\left(\boldsymbol{A}^{(0)}_{encoded}\right)$$

DeePathNet contains two layers of Transformer block, therefore:

$$\boldsymbol{A}^{(2)}_{encoded} = Transformer\left(\boldsymbol{A}^{(1)}_{encoded}\right)$$

Finally, in Step 3, DeePathNet uses a MLP to map $\boldsymbol{A}^{(2)}_{encoded}$ to the final prediction. The output dimension of a MLP depends on the prediction task. For drug response prediction, the

number of output dimensions is equal to the number of drugs, and for cancer type and subtype classification, the number of output dimensions is equal to the number of cancer types and subtypes.

Model training. All methods were trained with default hyperparameters for both regression and classification tasks. The default hyperparameters of DeePathNet and optimiser used can be found in the GitHub repository. Default hyperparameters were used for random forest, elastic net, PCA (top 200 PCs) and *k*-NN (*k* = 5) and details can be found in the official API of scikit-learn (v1.0.2). Default hyperparameters were also set for mixOmics and moCluster, and details can be found in their original publications. To train DeePathNet for regression, mean squared error (MSE) loss was computed between the predicted and actual IC_{50}. For classification, we computed the cross-entropy (CE) loss to train DeePathNet.

Evaluation metrics. For regression, R^2, MAE and Pearson's *r* were used to evaluate the performance and they are defined as follows:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

$$MAE = \frac{\sum_{i=1}^{n} |(y_i - \hat{y}_i)|}{n}$$

$$Pearson's\ r = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^{n} \left(\hat{y}_i - \bar{\hat{y}}\right)^2}}$$

For a given drug, y_i represents the actual IC_{50} of cell line i, $\hat{y}_i$ represents the predicted IC_{50} value of cell line i, $\bar{y}$ represents the mean value of all actual IC_{50} values, $\bar{\hat{y}}$ represents the mean value of all predicted IC_{50} values, and n represents the total number of cell lines. For classification, multiple metrics were used to evaluate the predictive performance of DeePathNet and other models, including accuracy, macro-average F_1-score, precision, recall, AUROC, AUPRC and stability. Let TP, TN, FP, FN represent true positive, true negative, false positive and false negative prediction. Accuracy is defined as $\frac{TP+TN}{TP+TN+FP+FN}$. Precision is defined as $\frac{TP}{TP+FP}$. Recall is defined as $\frac{TP}{TP+FN}$,. Then the F_1-score is calculated as the harmonic mean of the precision and recall and defined as $\frac{2 \cdot Precision \cdot Recall}{Precision+Recall}$. The macro-average F_1-score

is calculated by computing the arithmetic mean of F_1-scores from all the cancer types or subtypes. The ROC curve is created by plotting the recall and false positive rate ($\frac{FP}{FP+TN}$) at various thresholds. AUROC is calculated as the area under the ROC curve. The precision and recall (PR) curve is created by plotting the precision and recall at various thresholds, and the AUPRC is calculated as the area under the PR curve. The stability is measured by the standard deviation.

AUTHOR CONTRIBUTIONS

Z.C. and Q.Z. conceived and designed the experiments and wrote the manuscript. Z.C. analysed the data and developed the method. R.C.P., A.A., P.J.R. and R.R.R. contributed to the manuscript.

ACKNOWLEDGMENTS

ProCan® is supported by the Australian Cancer Research Foundation, Cancer Institute New South Wales (NSW) (2017/TPG001,REG171150), NSW Ministry of Health (CMP-01), The University of Sydney, Cancer Council NSW (IG 18-01), Ian Potter Foundation, the Medical Research Futures Fund (MRFF-PD), National Health and Medical Research Council (NHMRC) of Australia European Union grant (GNT1170739, a companion grant to support the European Commission's Horizon 2020 Program, H2020-SC1-DTH-2018-1, 'iPC-individualized Paediatric Cure' [ref. 826121]), and National Breast Cancer Foundation (IIRS-18-164). Work at ProCan® is done under the auspices of a Memorandum of Understanding between Children's Medical Research Institute and the U.S. National Cancer Institute's International Cancer Proteogenome Consortium (ICPC), that encourages cooperation among institutions and nations in proteogenomic cancer research in which datasets are made available to the public. R.C.P. is supported by the NHMRC of Australia (GNT1138536 and GNT2000855). Some figures were created with BioRender.com.

CONFLICT OF INTEREST

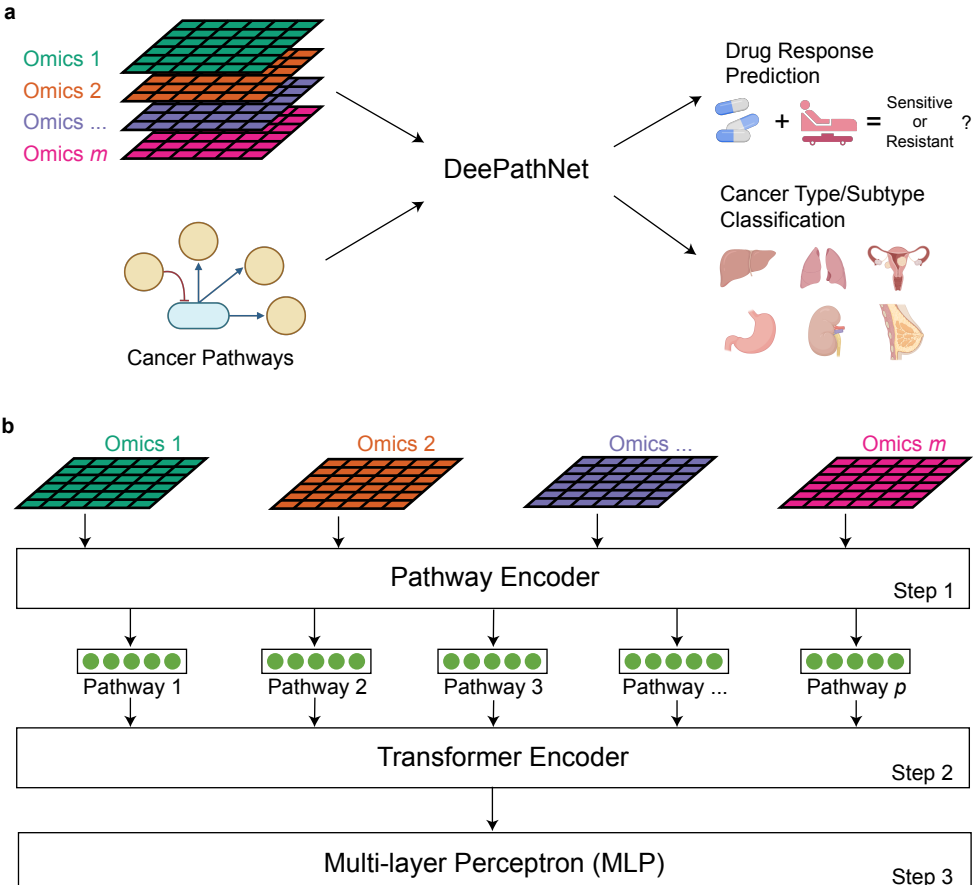
The authors declare no conflict of interest.

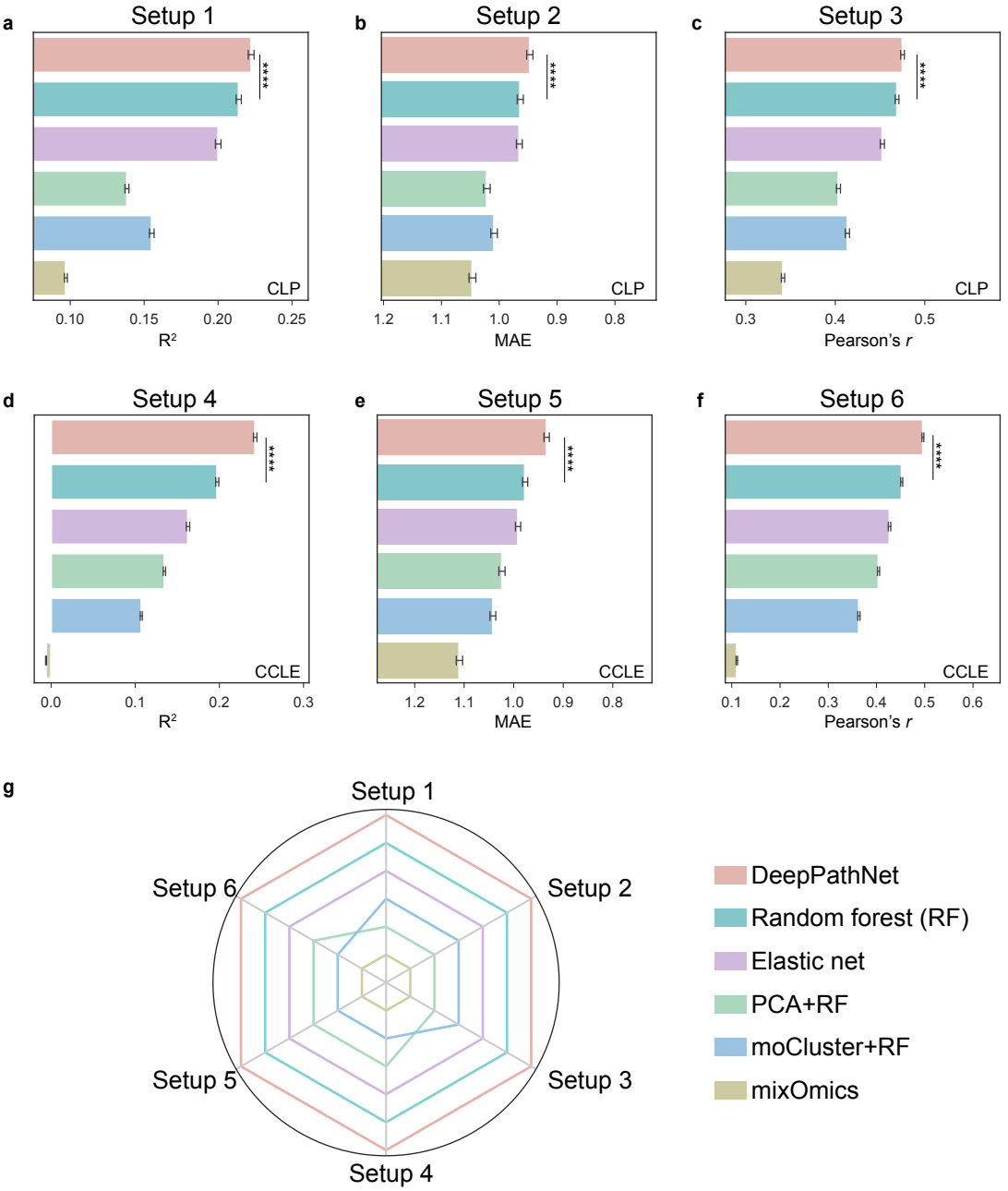
References

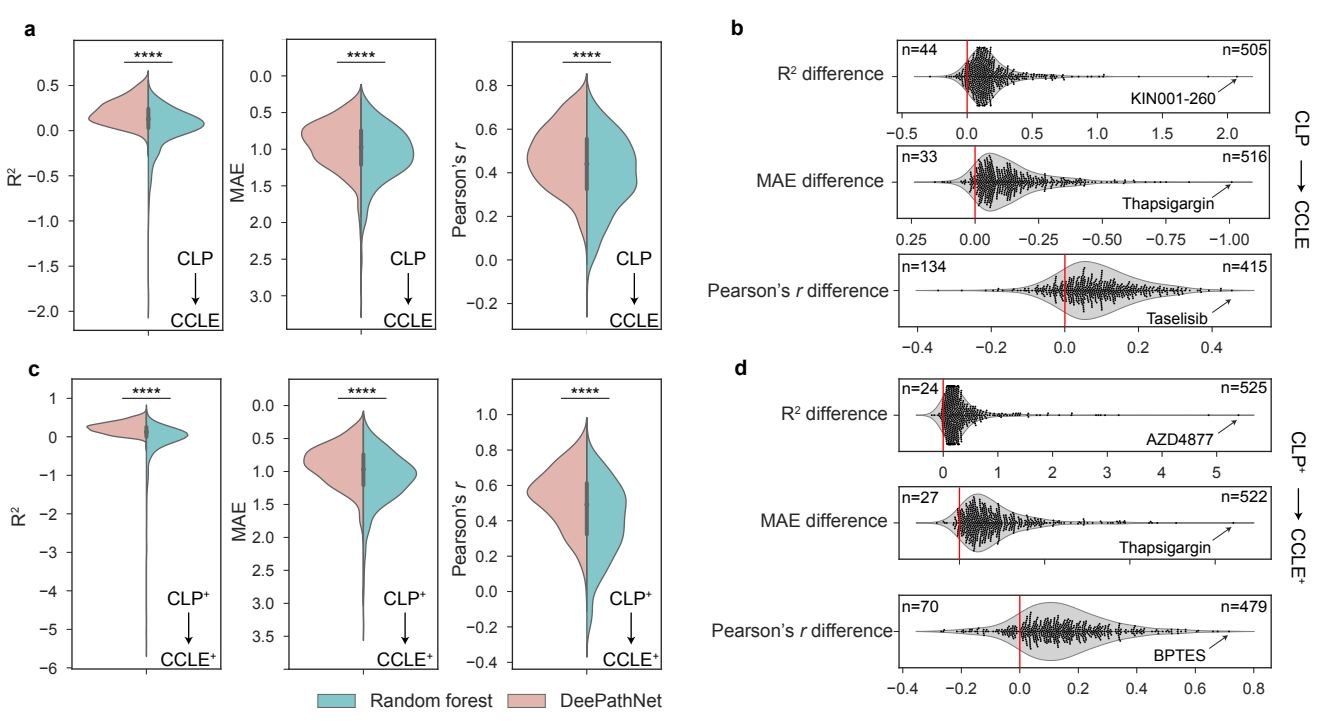
1. Mani, D. R. *et al.* Cancer proteogenomics: current impact and future prospects. *Nat. Rev. Cancer* **22**, 298–313 (2022).
2. Tsimberidou, A. M., Fountzilas, E., Nikanjam, M. & Kurzrock, R. Review of precision cancer medicine: Evolution of the treatment paradigm. *Cancer Treat. Rev.* **86**, 102019 (2020).
3. Gonçalves, E. *et al.* Pan-cancer proteomic map of 949 human cell lines. *Cancer Cell* **40**, 835-849.e8 (2022).
4. Tully, B. *et al.* Addressing the Challenges of High-Throughput Cancer Tissue Proteomics for Clinical Application: ProCan. *Proteomics* **19**, e1900109 (2019).
5. Rodriguez, H., Zenklusen, J. C., Staudt, L. M., Doroshow, J. H. & Lowy, D. R. The next horizon in precision oncology: Proteogenomics to inform cancer diagnosis and treatment. *Cell* **184**, 1661–1670 (2021).
6. Reel, P. S., Reel, S., Pearson, E., Trucco, E. & Jefferson, E. Using machine learning approaches for multi-omics data analysis: A review. *Biotechnol. Adv.* **49**, 107739 (2021).
7. Picard, M., Scott-Boyer, M.-P., Bodein, A., Périn, O. & Droit, A. Integration strategies of multi-omics data for machine learning analysis. *Comput. Struct. Biotechnol. J.* **19**, 3735–3746 (2021).
8. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: An R package for ’omics feature selection and multiple data integration. *PLoS Comput. Biol.* **13**, e1005752 (2017).
9. Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinform. Biol. Insights* **14**, 1177932219899051 (2020).
10. Tarazona, S., Arzalluz-Luque, A. & Conesa, A. Undisclosed, unmet and neglected challenges in multi-omics studies. *Nature Computational Science* **1**, 395–402 (2021).
11. Cai, Z., Poulos, R. C., Liu, J. & Zhong, Q. Machine learning for multi-omics data integration in cancer. *iScience* **25**, 103798 (2022).
12. Meng, C., Helm, D., Frejno, M. & Kuster, B. moCluster: Identifying Joint Patterns Across Multiple Omics Data Sets. *J. Proteome Res.* **15**, 755–765 (2016).
13. Shen, R., Olshen, A. B. & Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25**, 2906–2912 (2009).
14. Mo, Q. *et al.* A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics* **19**, 71–86 (2018).
15. Singh, A. *et al.* DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* **35**, 3055–3062 (2019).
16. Karlebach, G. & Shamir, R. Modelling and analysis of gene regulatory networks. *Nat. Rev. Mol. Cell Biol.* **9**, 770–780 (2008).
17. Kang, M., Ko, E. & Mersha, T. B. A roadmap for multi-omics data integration using deep learning. *Brief. Bioinform.* **23**, (2022).
18. Ma, J. *et al.* Using deep learning to model the hierarchical structure and function of a cell. *Nat. Methods* **15**, 290–298 (2018).
19. Kuenzi, B. M. *et al.* Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells. *Cancer Cell* **38**, 672-684.e6 (2020).

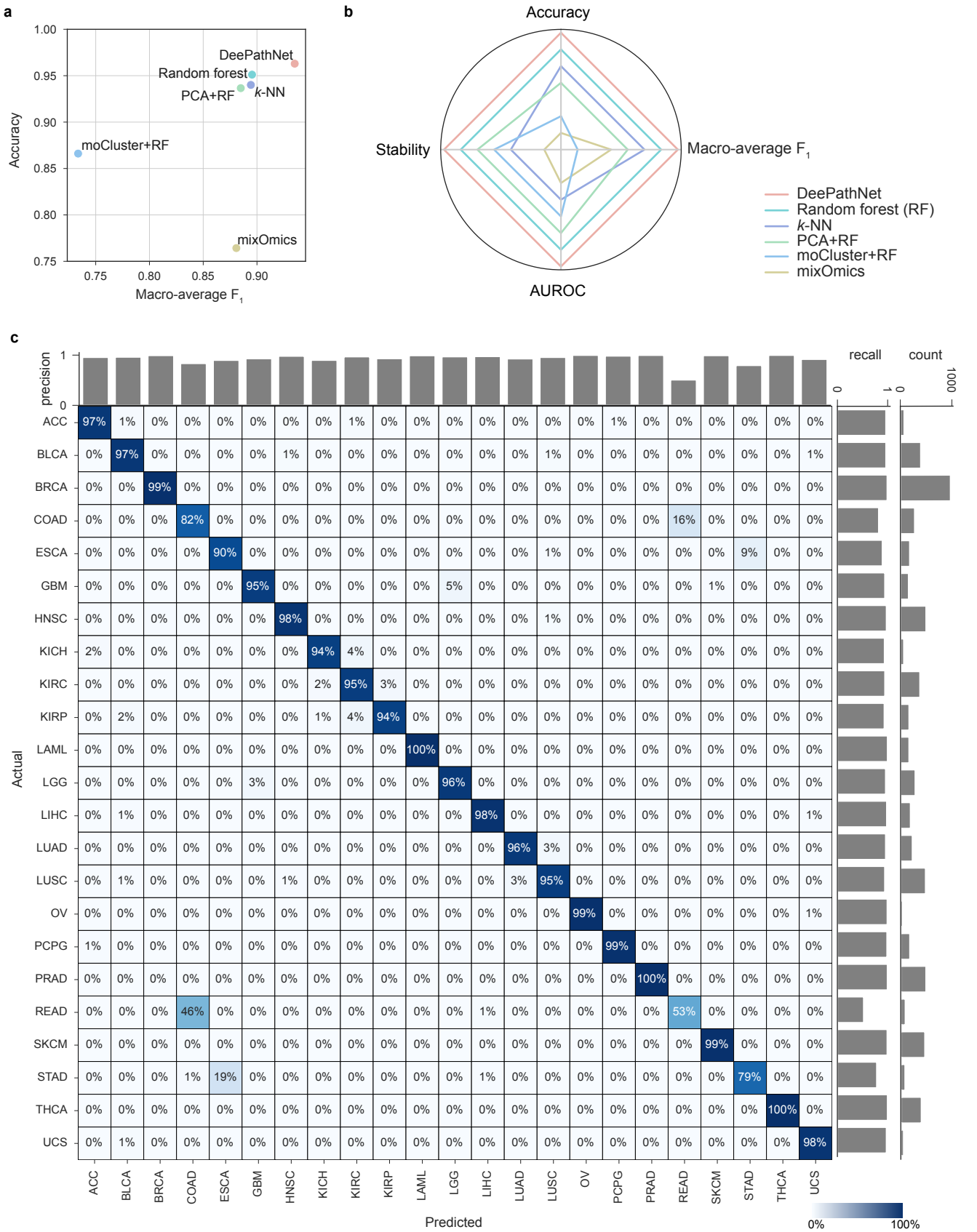
20. Schulte-Sasse, R., Budach, S., Hnisz, D. & Marsico, A. Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. *Nature Machine Intelligence* **3**, 513–526 (2021).
21. Wu, Z. *et al.* A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **32**, 4–24 (2021).
22. Kuenzi, B. M. & Ideker, T. A census of pathway maps in cancer systems biology. *Nat. Rev. Cancer* **20**, 233–246 (2020).
23. Vaswani, A. *et al.* Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, (2017).
24. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
25. Ho, T. K. Random decision forests. in *Proceedings of 3rd International Conference on Document Analysis and Recognition* (IEEE Comput. Soc. Press, 2002). doi:10.1109/icdar.1995.598994.
26. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.* **67**, 301–320 (2005).
27. Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740–754 (2016).
28. Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).
29. Nusinow, D. P. *et al.* Quantitative Proteomics of the Cancer Cell Line Encyclopedia. *Cell* **180**, 387-402.e16 (2020).
30. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
31. Fix, E. & Hodges, J. L. Discriminatory analysis. Nonparametric discrimination: Consistency properties. *Int. Stat. Rev.* **57**, 238 (1989).
32. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
33. Akiyama, H., Tsurumaru, M., Udagawa, H. & Kajiyama, Y. Esophageal cancer. *Curr. Probl. Surg.* **34**, 765–834 (1997).
34. Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160–1167 (2009).
35. Dai, X. *et al.* Breast cancer intrinsic subtype classification, clinical use and future trends. *Am. J. Cancer Res.* **5**, 2929–2943 (2015).
36. Raj-Kumar, P.-K. *et al.* PCA-PAM50 improves consistency between breast cancer intrinsic and clinical subtyping reclassifying a subset of luminal A tumors as luminal B. *Sci. Rep.* **9**, 7956 (2019).
37. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in *Advances in Neural Information Processing Systems 30* (eds. Guyon, I. et al.) 4765–4774 (Curran Associates, Inc., 2017).
38. Bach, S. *et al.* On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS One* **10**, e0130140 (2015).
39. Lu, X.-F. *et al.* FoxM1 is a promising candidate target in the treatment of breast cancer. *Oncotarget* **9**, 842–852 (2018).
40. Lee, J.-J. *et al.* Expression of FOXM1 and related proteins in breast cancer molecular subtypes. *Int. J. Exp. Pathol.* **97**, 170–177 (2016).
41. Marchesin, V. *et al.* ARF6-JIP3/4 regulate endosomal tubules for MT1-MMP exocytosis in cancer invasion. *J. Cell Biol.* **211**, 339–358 (2015).
42. Acar, A., Simões, B. M., Clarke, R. B. & Brennan, K. A Role for Notch Signalling in Breast Cancer and Endocrine Resistance. *Stem Cells Int.* **2016**, 2498764 (2016).

43. Satpathy, S. *et al.* A proteogenomic portrait of lung squamous cell carcinoma. *Cell* **184**, 4348-4371.e40 (2021).
44. Zhang, Y., Chen, F., Chandrashekar, D. S., Varambally, S. & Creighton, C. J. Proteogenomic characterization of 2002 human cancers reveals pan-cancer molecular subtypes and associated pathways. *Nat. Commun.* **13**, 2669 (2022).
45. Esteva, A. *et al.* A guide to deep learning in healthcare. *Nat. Med.* **25**, 24–29 (2019).
46. Wei, L. *et al.* TCGA-assembler 2: software pipeline for retrieval and processing of TCGA/CPTAC data. *Bioinformatics* **34**, 1615–1617 (2018).
47. Krug, K. *et al.* Proteogenomic Landscape of Breast Cancer Tumorigenesis and Targeted Therapy. *Cell* **183**, 1436-1456.e31 (2020).
48. Lin, T., Wang, Y., Liu, X. & Qiu, X. A survey of transformers. *arXiv [cs.LG]* (2021).

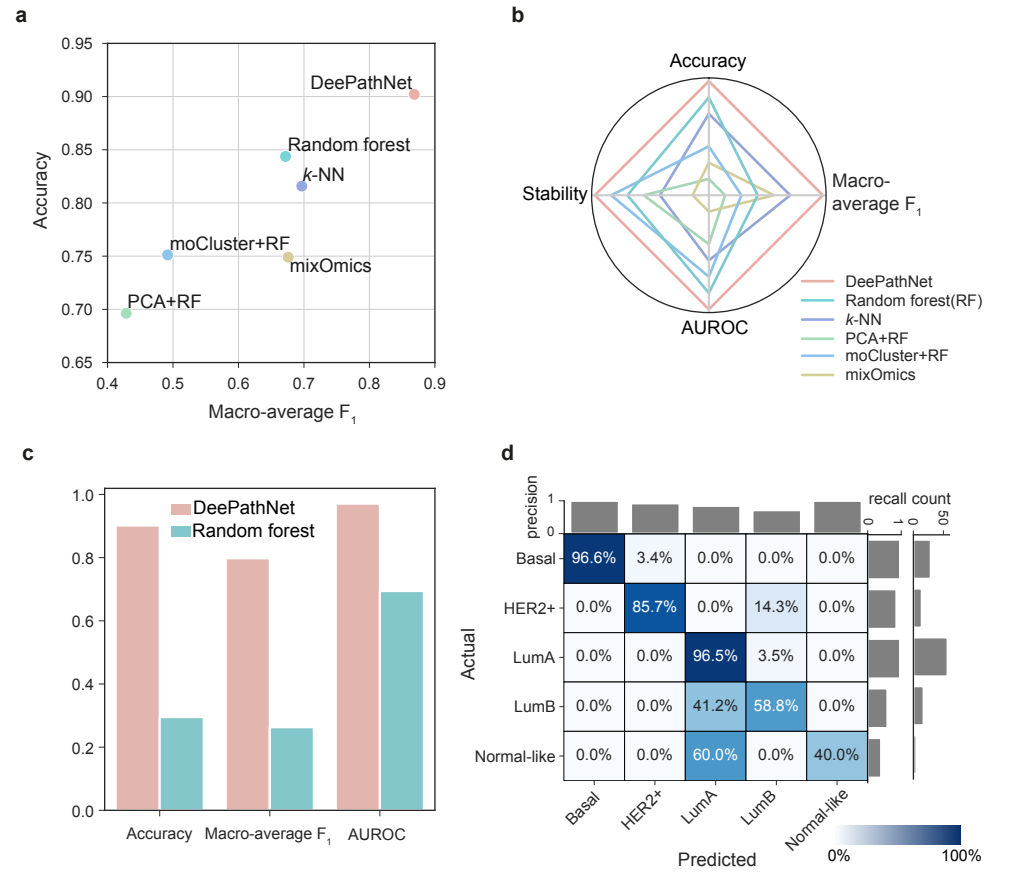

a
Omics 1
Omics 2
Omics ...
Omics m
Cancer Pathways
DeePathNet
Drug Response Prediction
Sensitive or Resistant ?
Cancer Type/Subtype Classification
b
Omics 1
Omics 2
Omics ...
Omics m
Pathway Encoder
Step 1
Pathway 1
Pathway 2
Pathway 3
Pathway ...
Pathway p
Transformer Encoder
Step 2
Multi-layer Perceptron (MLP)
Step 3

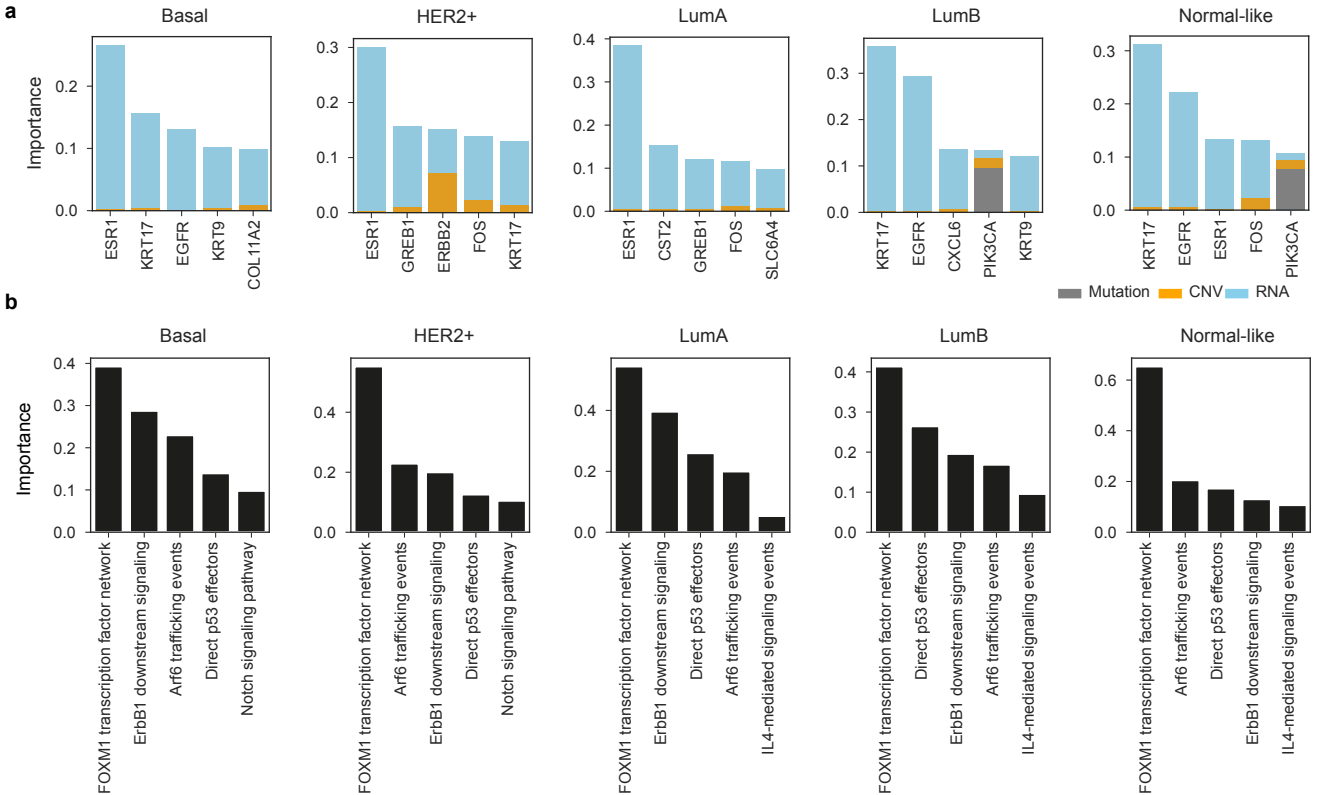

a
Setup 1
****
CLP
0.10
0.15
0.20
0.25
R²
b
Setup 2
****
CLP
1.2
1.1
1.0
0.9
0.8
MAE
c
Setup 3
****
CLP
0.3
0.4
0.5
Pearson's r
d
Setup 4
****
CCLE
0.0
0.1
0.2
0.3
R²
e
Setup 5
****
CCLE
1.2
1.1
1.0
0.9
0.8
MAE
f
Setup 6
****
CCLE
0.1
0.2
0.3
0.4
0.5
0.6
Pearson's r
g
Setup 1
Setup 2
Setup 3
Setup 4
Setup 5
Setup 6
DeepPathNet
Random forest (RF)
Elastic net
PCA+RF
moCluster+RF
mixOmics

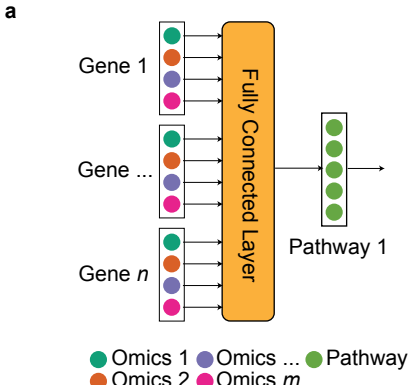

a
****
R²
MAE
Pearson's r
CLP
CCLE
b
R² difference
MAE difference
Pearson's r difference
n=44
n=505
KIN001-260
n=33
n=516
Thapsigargin
n=134
n=415
Taselisib
CLP → CCLE
c
CLP⁺
CCLE⁺
d
n=24
n=525
AZD4877
n=27
n=522
Thapsigargin
n=70
n=479
BPTES
CLP⁺ → CCLE⁺
Random forest
DeePathNet

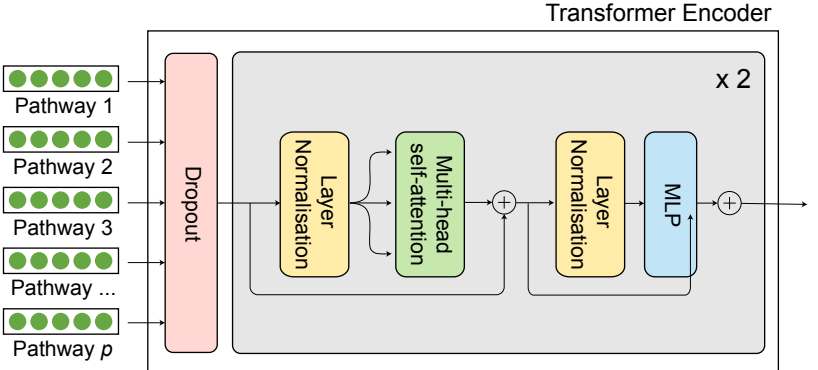


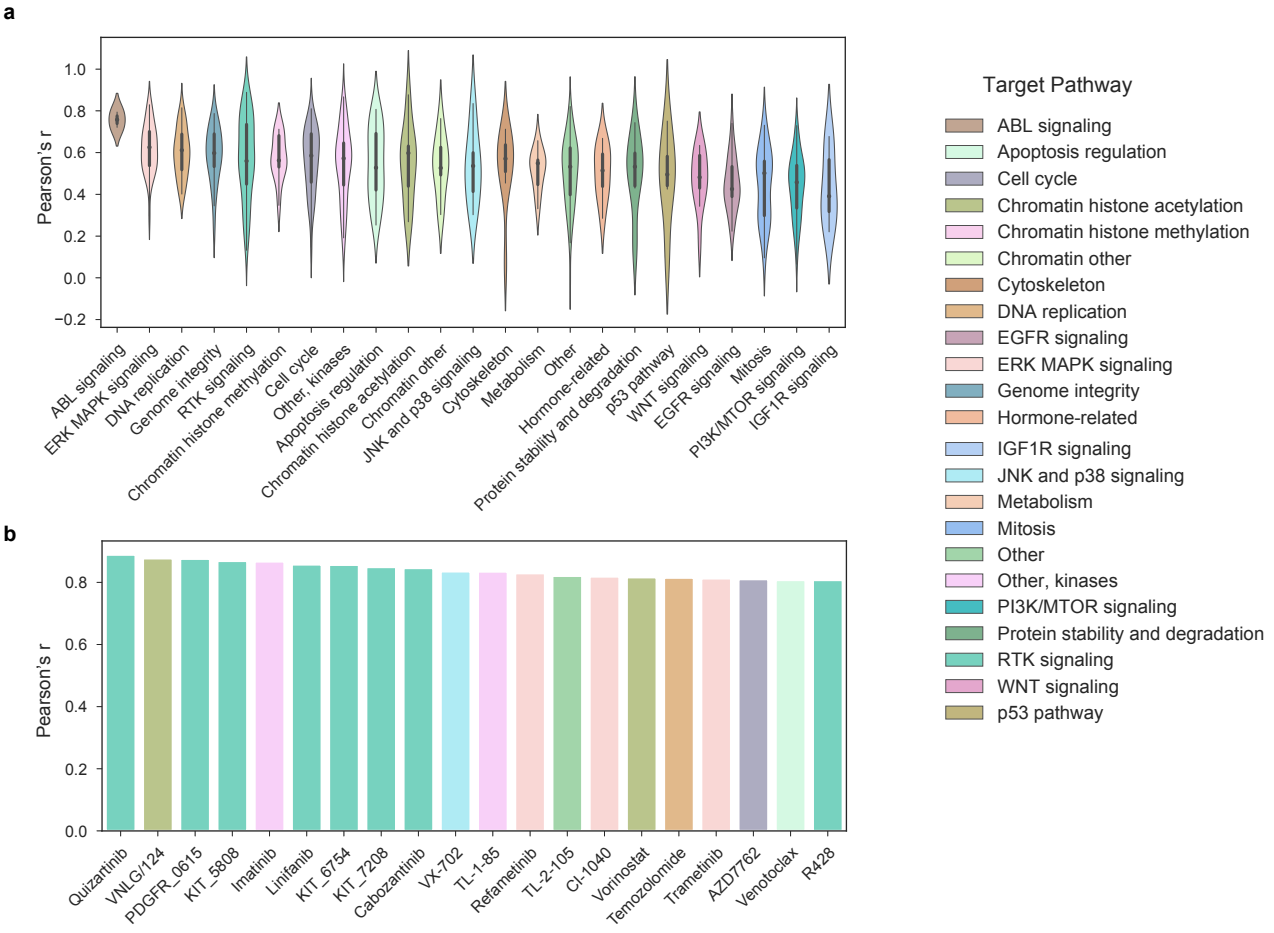
ACC – Adrenocortical carcinoma
BLCA – Bladder urothelial carcinoma
BRCA – Breast invasive carcinoma
COAD – Colon adenocarcinoma
ESCA – Esophageal carcinoma
GBM – Glioblastoma multiforme
HNSC – Head and neck squamous cell carcinoma
KICH – Kidney chromophobe
KIRC – Kidney renal clear cell carcinoma
KIRP – Kidney renal papillary cell carcinoma
LAML – Acute myeloid leukemia
LGG – Brain lower grade glioma
LIHC – Liver hepatocellular carcinoma
LUAD – Lung adenocarcinoma
LUSC – Lung squamous cell carcinoma
OV – Ovarian serous cystadenocarcinoma
PCPG – Pheochromocytoma and paraganglioma
PRAD – Prostate adenocarcinoma
READ – Rectum adenocarcinoma
SKCM – Skin cutaneous melanoma
STAD – Stomach adenocarcinoma
THCA – Thyroid carcinoma
UCS – Uterine carcinosarcoma

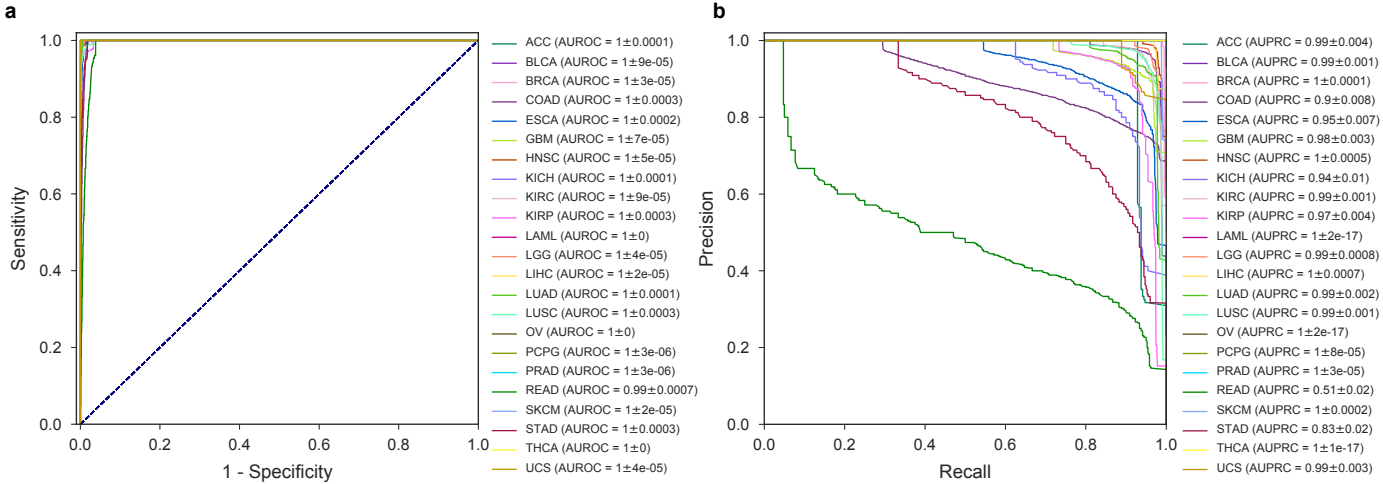

a
0.95
0.90
0.85
0.80
0.75
0.70
0.65
Accuracy
DeePathNet
Random forest
k-NN
moCluster+RF
mixOmics
PCA+RF
0.4
0.5
0.6
0.7
0.8
0.9
Macro-average F1
b
Accuracy
Stability
Macro-average F1
AUROC
DeePathNet
Random forest(RF)
k-NN
PCA+RF
moCluster+RF
mixOmics
c
1.0
0.8
0.6
0.4
0.2
0.0
DeePathNet
Random forest
Accuracy
Macro-average F1
AUROC
d
precision
1
0
recall count
0
1
0
50
Actual
Basal 96.6% 3.4% 0.0% 0.0% 0.0%
HER2+ 0.0% 85.7% 0.0% 14.3% 0.0%
LumA 0.0% 0.0% 96.5% 3.5% 0.0%
LumB 0.0% 0.0% 41.2% 58.8% 0.0%
Normal-like 0.0% 0.0% 60.0% 0.0% 40.0%
Basal
HER2+
LumA
LumB
Normal-like
Predicted
0%
100%

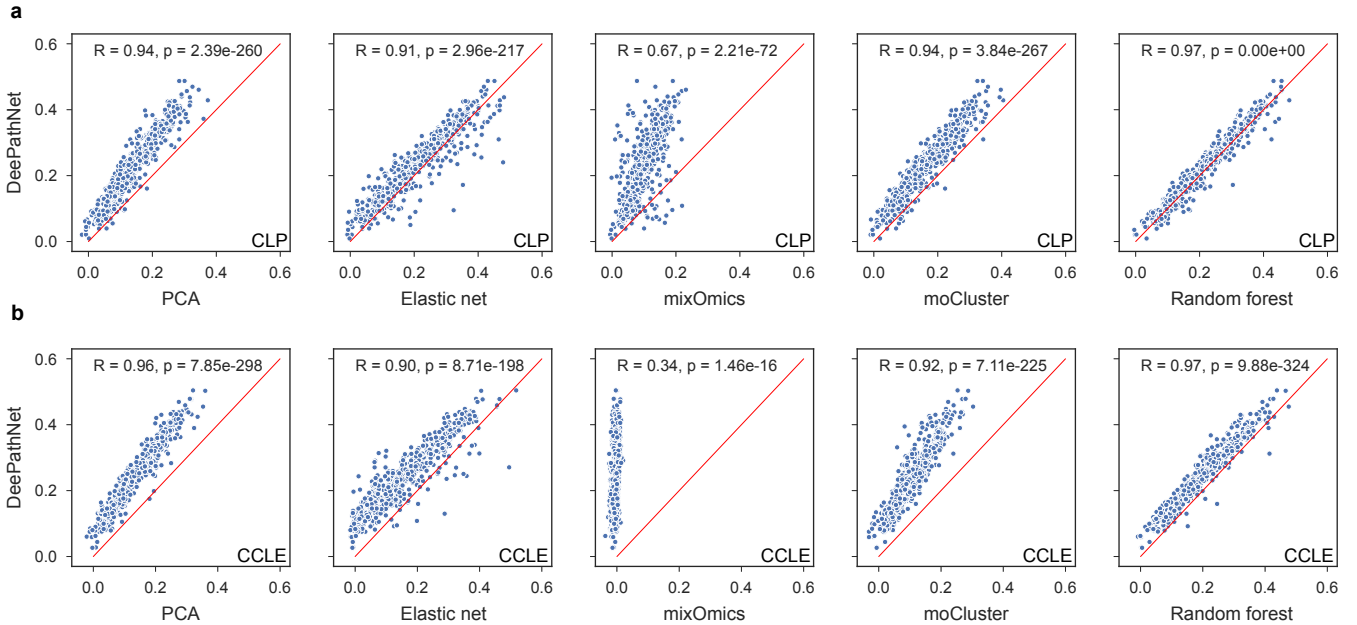

a
Basal
HER2+
LumA
LumB
Normal-like
Importance
ESR1
KRT17
EGFR
KRT9
COL11A2
ESR1
GREB1
ERBB2
FOS
KRT17
ESR1
CST2
GREB1
FOS
SLC6A4
KRT17
EGFR
CXCL6
PIK3CA
KRT9
KRT17
EGFR
ESR1
FOS
PIK3CA
Mutation
CNV
RNA
b
Basal
HER2+
LumA
LumB
Normal-like
Importance
FOXM1 transcription factor network
ErbB1 downstream signaling
Arf6 trafficking events
Direct p53 effectors
Notch signaling pathway
FOXM1 transcription factor network
Arf6 trafficking events
ErbB1 downstream signaling
Direct p53 effectors
Notch signaling pathway
FOXM1 transcription factor network
ErbB1 downstream signaling
Direct p53 effectors
Arf6 trafficking events
IL4-mediated signaling events
FOXM1 transcription factor network
Direct p53 effectors
ErbB1 downstream signaling
Arf6 trafficking events
IL4-mediated signaling events
FOXM1 transcription factor network
Arf6 trafficking events
Direct p53 effectors
ErbB1 downstream signaling
IL4-mediated signaling events

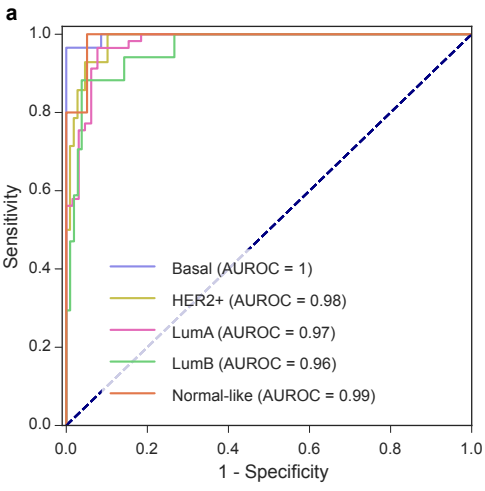

a
Gene 1
Gene ...
Gene n
Fully Connected Layer
Pathway 1
Omics 1
Omics 2
Omics ...
Omics m
Pathway

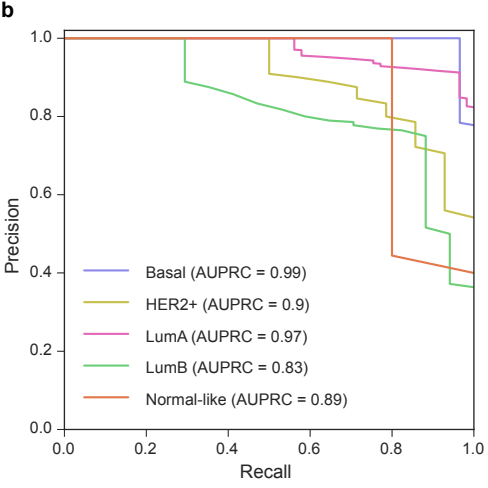

b
Transformer Encoder
Pathway 1
Pathway 2
Pathway 3
Pathway ...
Pathway p
Dropout
x 2
Layer Normalisation
Multi-head self-attention
Layer Normalisation
MLP


a
Pearson's r
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
ABL signaling
ERK MAPK signaling
DNA replication
Genome integrity
RTK signaling
Chromatin histone methylation
Cell cycle
Other, kinases
Apoptosis regulation
Chromatin histone acetylation
Chromatin other
JNK and p38 signaling
Cytoskeleton
Metabolism
Other
Hormone-related
Protein stability and degradation
p53 pathway
WNT signaling
EGFR signaling
Mitosis
PI3K/MTOR signaling
IGF1R signaling
b
Pearson's r
0.8
0.6
0.4
0.2
0.0
Quizartinib
VNLG/124
PDGFR_0615
KIT_5808
Imatinib
Linifanib
KIT_6754
KIT_7208
Cabozantinib
VX-702
TL-1-85
Refametinib
TL-2-105
CI-1040
Vorinostat
Temozolomide
Trametinib
AZD7762
Venotoclax
R428
Target Pathway
ABL signaling
Apoptosis regulation
Cell cycle
Chromatin histone acetylation
Chromatin histone methylation
Chromatin other
Cytoskeleton
DNA replication
EGFR signaling
ERK MAPK signaling
Genome integrity
Hormone-related
IGF1R signaling
JNK and p38 signaling
Metabolism
Mitosis
Other
Other, kinases
PI3K/MTOR signaling
Protein stability and degradation
RTK signaling
WNT signaling
p53 pathway


a
Sensitivity
1 - Specificity
0.0
0.2
0.4
0.6
0.8
1.0
ACC (AUROC = 1±0.0001)
BLCA (AUROC = 1±9e-05)
BRCA (AUROC = 1±3e-05)
COAD (AUROC = 1±0.0003)
ESCA (AUROC = 1±0.0002)
GBM (AUROC = 1±7e-05)
HNSC (AUROC = 1±5e-05)
KICH (AUROC = 1±0.0001)
KIRC (AUROC = 1±9e-05)
KIRP (AUROC = 1±0.0003)
LAML (AUROC = 1±0)
LGG (AUROC = 1±4e-05)
LIHC (AUROC = 1±2e-05)
LUAD (AUROC = 1±0.0001)
LUSC (AUROC = 1±0.0003)
OV (AUROC = 1±0)
PCPG (AUROC = 1±3e-06)
PRAD (AUROC = 1±3e-06)
READ (AUROC = 0.99±0.0007)
SKCM (AUROC = 1±2e-05)
STAD (AUROC = 1±0.0003)
THCA (AUROC = 1±0)
UCS (AUROC = 1±4e-05)
b
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
ACC (AUPRC = 0.99±0.004)
BLCA (AUPRC = 0.99±0.001)
BRCA (AUPRC = 1±0.0001)
COAD (AUPRC = 0.9±0.008)
ESCA (AUPRC = 0.95±0.007)
GBM (AUPRC = 0.98±0.003)
HNSC (AUPRC = 1±0.0005)
KICH (AUPRC = 0.94±0.01)
KIRC (AUPRC = 0.99±0.001)
KIRP (AUPRC = 0.97±0.004)
LAML (AUPRC = 1±2e-17)
LGG (AUPRC = 0.99±0.0008)
LIHC (AUPRC = 1±0.0007)
LUAD (AUPRC = 0.99±0.002)
LUSC (AUPRC = 0.99±0.001)
OV (AUPRC = 1±2e-17)
PCPG (AUPRC = 1±8e-05)
PRAD (AUPRC = 1±3e-05)
READ (AUPRC = 0.51±0.02)
SKCM (AUPRC = 1±0.0002)
STAD (AUPRC = 0.83±0.02)
THCA (AUPRC = 1±1e-17)
UCS (AUPRC = 0.99±0.003)


a
R = 0.94, p = 2.39e-260
R = 0.91, p = 2.96e-217
R = 0.67, p = 2.21e-72
R = 0.94, p = 3.84e-267
R = 0.97, p = 0.00e+00
CLP
DeePathNet
PCA
Elastic net
mixOmics
moCluster
Random forest
b
R = 0.96, p = 7.85e-298
R = 0.90, p = 8.71e-198
R = 0.34, p = 1.46e-16
R = 0.92, p = 7.11e-225
R = 0.97, p = 9.88e-324
CCLE
0.0
0.2
0.4
0.6


a
Sensitivity
1 - Specificity
0.0
0.2
0.4
0.6
0.8
1.0
Basal (AUROC = 1)
HER2+ (AUROC = 0.98)
LumA (AUROC = 0.97)
LumB (AUROC = 0.96)
Normal-like (AUROC = 0.99)


b
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
Basal (AUPRC = 0.99)
HER2+ (AUPRC = 0.9)
LumA (AUPRC = 0.97)
LumB (AUPRC = 0.83)
Normal-like (AUPRC = 0.89)