nature communications

Article

https://doi.org/10.1038/s41467-025-64506-8

UNICORN: Towards universal cellular expression prediction with a multi-task learning framework

Received: 14 January 2025

Accepted: 18 September 2025

Published online: 27 October 2025

Check for updates

Tianyu Liu[1,2], Tinglin Huang[3], Lijun Wang[2], Yingxin Lin[2], Rex Ying[3] & Hongyu Zhao[1,2]

Sequence-to-function analysis is a challenging task in human genetics, especially in predicting cell-type-specific multi-omic phenotypes from biological sequences such as individualized gene expression. Here, we present UNICORN, a computational method with improved prediction performances than the existing methods. UNICORN takes the embeddings from biological sequences as well as external knowledge from pre-trained foundation models as inputs and optimizes the predictor with carefully-designed loss functions. We demonstrate that UNICORN outperforms the existing methods in both gene expression prediction and multi-omic phenotype prediction at the cellular level and the cell-type level, and it can also generate uncertainty scores of the predictions. Moreover, UNICORN is able to link personalized gene expression profiles with corresponding genome information. Finally, we show that UNICORN is capable of characterizing complex biological systems for different disease states or perturbations. Overall, embeddings from foundation models can facilitate the understanding of the role of biological sequences in the prediction task, and incorporating multi-omic information can enhance prediction performances.

Modeling DNA and the Central Dogma is a challenging but essential task to biological research[1–7]. One key question is how to predict the expression of products from sequences (e.g., DNA sequences[8] and amino acid sequences[9,10]). A good predictor can help us uncover the relationship between genetic information and cellular functions, which cannot be directly achieved by using previous sequence-to-expression models due to the difference of training data[11]. With the accumulation of single-cell sequencing data[12–14], it is possible to impute and analyze cell-specific or cell-type-specific functional information based on sequence models, an exciting direction to link genomics and transcriptomics/proteomics with a unified framework to advance genetic research.

Single-cell sequencing (including scRNA-seq[12,13], scATAC-seq[15], and CITE-seq[16]) generates different types of measurements at the single-cell resolution. The knowledge gained from downstream analysis, such as cell-type or cell-cycle information, may raise biological questions to explore. Although single-cell data allow us to capture the omic information at the single-cell resolution, there is large noise in these data, which presents a major obstacle for data analysis and interpretation[17]. To tackle this problem,[18] considers aggregating single-cell data into pseudo-bulk data using metadata information of cells. Such an approach is widely used in inferring cell-type-specific expression quantitative trait loci (eQTLs)[19] or evaluations for deconvolution[20]. However, the effectiveness of pseudo-bulk settings in sequence-to-expression tasks has not yet been discussed in detail. Meanwhile, these predictors may also have biases for certain cell types or diseases[21,22], which motivates us to explore the reasons behind such performance differences.

Existing methods for gene expression prediction based on DNA sequences relied on either CAGE-seq results (e.g., Basenji2[23], Enformer[21])

[1]Interdepartmental Program in Computational Biology & Bioinformatics, Yale University, New Haven, CT, USA. [2]Department of Biostatistics, Yale University, New Haven, CT, USA. [3]Department of Computer Science, Yale University, New Haven, CT, USA. e-mail: hongyu.zhao@yale.edu

or bulkRNA-seq results (e.g., Borzoi[24]) as targets for training. Most of them were developed before significant progress was made in single-cell atlas, and the cost of bulk-level data is less than sequencing single-cell data[25]. Nvwa[26] attempted to predict whether a gene will express in cells using single-cell transcriptomics, but non-human data were considered, and classification is not a proper problem setting for gene expression prediction because it is important to evaluate the expression levels. Recently, seq2cells[11] built on the pre-trained Enformer model to extract embeddings from DNA sequences to predict gene expression across all the cells, whereas scooby[27] used a similar idea for single-cell multi-omic data. Moreover, researchers are also interested in exploring the modeling of DNA language based on a self-supervised learning[28] (SSL) approach (e.g., HyenaDNA[2], DNABERT[29], Evo[30]). However, these methods have specific limitations. First, models similar to Enformer did not consider enough cell-type-specific information, and thus, the heterogeneity of gene expression at the cell level cannot be fully modeled. Second, seq2cells did not discuss the choices of models for generating sequence embeddings, nor did it discuss the selection process of targets for prediction. The prediction of single-cell gene expression is also not practical because of the experimental noise. Third, SSL-based DNA language models did not always show better performance in the evaluations of downstream tasks[31], and few are applied for cell-type-level gene expression prediction. Finally, there are some technical limitations in constructing a sequence-to-expression model, such as how to model meaningful sequence context and how to reflect the individual-level effects in the prediction results. Therefore, we need a flexible and transparent framework to answer these questions.

In this manuscript, we introduce an advanced transfer-learning-based framework for predicting multi-omic expression levels from their corresponding sequence information, called a **UNI**versal **C**ell expressi**O**n p**R**ediction**N** framework (UNICORN). We utilize a general genetic language model[32] plus a general large language model (LLM)[33] as an option to generate the embeddings of sequences and predict the corresponding expression levels (also defined as phenotypes in our context[34]) of such sequences (genes, peaks, or proteins) in different cell types at the pseudo-bulk level. Our framework is explainable with the help of uncertainty prediction for the expression levels of different sequences and multi-task-learnable with the help of different types of omic data used for training and testing. UNICORN is a powerful framework to understand the relationships between 1. biological sequence and cell-type-specific effects, 2. biological sequence and individual genetic effects, and 3. biological sequence and disease/perturbation-specific effects.

Results

Overview of UNICORN

Our model predicts different types of omic data based on their corresponding biological sequence information. We first infer the embeddings of target biological sequences with corresponding language models (genomic language models (gLMs), protein language models (PLMs), or LLMs) and save the embeddings **e** in a separate file. We then construct the training and testing datasets based on the target omic type and train a non-linear predictor f_{θ_1} to model the relationship between sequence and expression information by predicting the expression-level information from the sequence-level information. Meanwhile, we have an uncertainty estimator to quantify the uncertainty of each prediction. The difference between UNICORN and other baseline models is summarized in Supplementary Fig. 1, which demonstrates the unique contributions of our model design. Our metrics include the correlation and distance between the observed and the predicted expression levels, and our model is capable of multiple downstream applications, including the analysis of genetic effects at the cell-type levels, individual levels, disease-specific levels, and perturbation-specific levels. The landscape of UNICORN is summarized in Fig. 1a–c. Details of UNICORN are presented in the Methods section.

Exploring the capacity of UNICORN for expression prediction

Here, we explore the performance of UNICORN towards predicting gene expression levels at the single-cell or cell-type level. This task is the basic component of our expression prediction problem and is helpful for us in analyzing the capacity of sequence-based predictors. To predict the expression levels of each cell, we referred to the original design of Enformer to collect sequence information of ~200 kb length

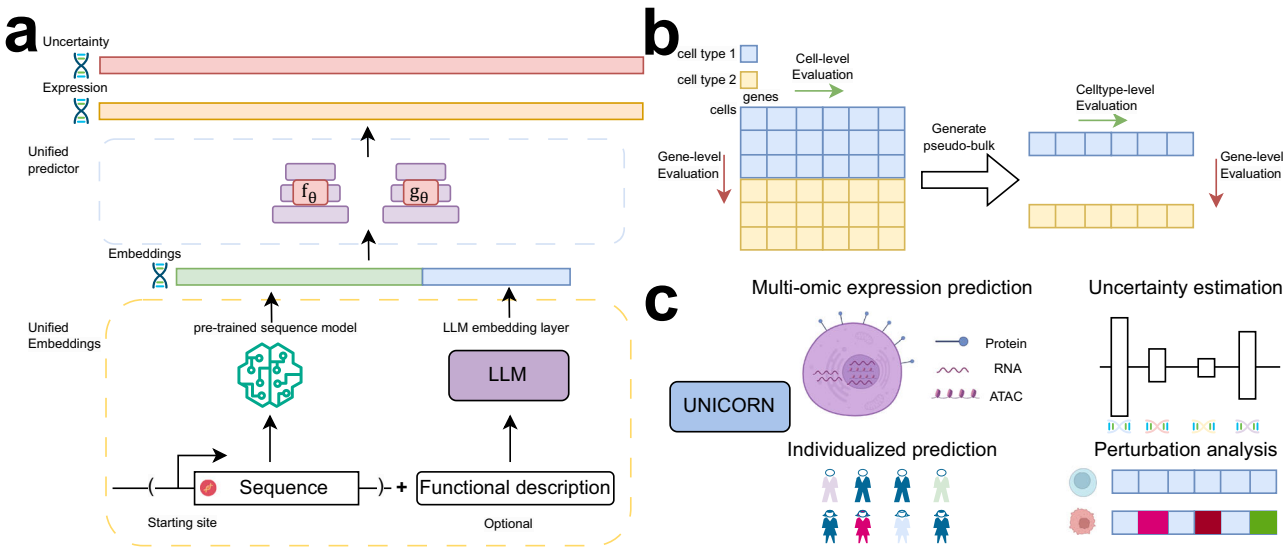


Fig. 1 | Workflow and functions of UNICORN. a The architecture and components of UNICORN. Our method accepts embeddings from multi-modal sources, predicts expression levels at the single-cell or cell-type resolutions, and estimates the uncertainty of model outputs. **b** Explanations of our evaluation framework. We consider evaluations from two perspectives (per-cell and per-gene) and two modes (single-cell level and pseudo-bulk level). **c** A summary of UNICORN's unique functions, including multi-omic expression prediction, uncertainty estimation, individualized prediction, and perturbation prediction. The figure is created by the authors using Draw.io[86]. The examples of cells are generated with BioRender.com (Created in BioRender. Liu, T. (2025) https://biorender.com/1lgm4br).

centered at the transcription start site (TSS) for each gene. We used the centered sequence to generate the gene embeddings as inputs and single-cell gene expression levels as outputs. We split a single-cell dataset into a training set (70%), a validation set (20%), and a testing set (10%) using gene names and chose mean squared error (MSE) and Pearson correlation coefficients (corr) as metrics for evaluations[11,35]. Our evaluations considered both genes and cells as dimensions and observed expression levels in the pseudo-bulk mode. For the models we evaluated, we chose baselines including Enformer (enformer), Borzoi (borzoi), seq2cells, UNICORN with gene embeddings from Enformer (UNICORN), gene embeddings from HyenaDNA (UNICORN_hyenadna), gene embeddings from Borzoi (UNICORN_borzoi), gene embeddings from Nucleotide Transformers[36] (UNICORN_nt), gene embeddings from DNABERT (UNICORN_dnabert), gene embeddings from Gene2vec[37] (UNICORN_gene2vec), gene embeddings from LLMs[38](UNICORN_genept), and gene embeddings from both Enformer and LLMs (UNICORN_comb). We considered two datasets for testing, including scRNA-seq data from Thymus (thymus) and scRNA-seq data from PBMC (pbmc). Details of our training settings, methods, and metrics can be found in the Methods section.

We present the prediction results at the single-cell resolution in Fig. 2a–d. All the methods are ranked by the mean values of their evaluation results in ascending order from left to right. Different figures correspond to different evaluation approaches. For the thymus dataset, UNICORN_comb had the highest correlations and lowest MSE. Figure 2a shows the comparisons based on correlations at the gene level, and we found that using more advanced models to encode sequence information (for example, HyenaDNA) did not yield better results compared with the default mode of UNICORN. Moreover, the other methods (UNICORN_gene2vec, UNICORN_comb, seq2cells, and UNICORN) did not show obvious differences by considering the distribution of correlations. However, the UNICORN and its combination mode had a smaller averaged MSE according to Fig. 2b, and thus a good approach to integrating gene embeddings from different modalities effectively can enhance prediction performances. Furthermore, we visualize the results with gene-level evaluations of the PBMC dataset in Fig. 2c, d. For this dataset, UNICORN and UNICORN_combine had high correlation scores, and UNICORN had the lowest MSE. We found that UNICORN with gene2vec, genept, and combined information as sources of gene embeddings had a larger averaged MSE, while full-parameter fine-tuning based on Enformer and Borzoi did not perform well for both datasets. The results of cell-level metrics of these two datasets are shown in Supplementary Fig. 2a, d), and UNICORN was always among the top two performers across different cell-level metrics. To further support the strength of our model design, we compared UNICORN with Decima[22], a model pre-trained with large-scale single-cell transcriptomic data for expression prediction, and found that our method outperformed Decima in the evaluation of average MSE and gene-level correlation, shown in Supplementary Fig. 3. Therefore, gene embeddings from sequences worked as a better input for predicting gene expression levels, compared with information directly from co-expression relationships or text mining. One possible reason to explain the contribution of gene embeddings for prediction is the positive effect of decoupling training[39]. In conclusion, our experimental results demonstrate that sequence information can serve as a better starting point for analyzing gene expression at the cellular level.

We also performed biology-oriented evaluation to demonstrate the superiority of UNICORN for predicting meaningful gene expression levels. According to Supplementary Fig.4a, b, UNICORN ranked in the top models for predicting the expression levels of cell-type marker genes in the Thymus dataset. Within the same dataset, we also computed clustering scores with predicted transcriptomic information at the cell level to examine the capacity of UNICORN and other models in distinguishing different cell types and generating meaningful cell clusters. According to our results in Supplementary Fig. 5, UNICORN still shows strong performance in representing cells from different cell types, supported by the high clustering scores by either metric-specific or average evaluations. Finally, we tested the proportion of extracted Gene Ontology (GO) pathways with the genes from different approaches with good prediction. To select important genes, we first removed the genes whose PCC is lower than the mean PCC or MSE is higher than the mean MSE from different methods, and run GO pathway analysis based on these gene sets. We then computed the proportion of significant pathways discovered in these gene sets. According to Supplementary Fig. 6, methods from the series of UNICORN still present leading performances in predicting gene expression levels with biological information.

We improved the training efficiency compared with seq2cells and Enformer, shown in Fig. 2g. We controlled the batch size as 1024 for different models to make a fair comparison, and UNICORN had the fastest training speed.

However, we note that the correlations measured at the single-cell level were generally low, especially for the PBMC dataset. Most methods did not have prediction results with an average correlation value higher than 0.05. This is largely due to the high noise in single-cell data measurements. Therefore, we aggregated single-cell data into pseudo-bulk data by cell types and visualized the results in Fig. 2e, f. Based on these two figures, correlations for both gene and cell levels increased, and UNICORN could better predict gene expressions at the cell type level with a significant improvement over Enformer based on the two-sided Mann-Whitney U test. Furthermore, UNICORN with combination embeddings as input data outperformed the default mode in general. Therefore, our conclusions for the single-cell resolution also hold for the cell-type resolution.

We further evaluated the capacity of predicting cell-type-level expression profiles with UNICORN based on two atlas-level scRNA-seq datasets, including the ROSMAP dataset from human brain tissue and the OneK1K dataset from human PBMC. We updated the baseline methods with different models, including SCOVER (scover)[40] and DanQ[41], which are based on the motif-captured design; LassoNet[42], which works as a feature-selection-driven method for multi-target regression; and GET[43], which is a foundation model for predicting gene expression levels from chromatin accessibility across different cell types. We retrained the previous three baseline methods with selected datasets for comparison. To compare GET, as it provided predicted gene expression profiles for different cell types, we directly took the predicted results for comparison. We utilized the same PCC and MSE as metrics to perform gene-level evaluation and cell-type-level evaluation. The results are shown in Supplementary Fig. 7a–d based on the OneK1K[44] dataset and Supplementary Fig. 8a–d based on the ROSMAP[45] dataset. According to these figures, UNICORN outperforms all selected baseline methods except GET for cell-level correlations. Our method also captured gene-level signals better than GET for both the ROSMAP and OneK1K datasets and was close to GET in capturing cell-type-level signals. Since GET was pre-trained with data from ATAC-seq, it had a high correlation with certain transcriptomic profiles[46,47].

We also found that more cells for training can improve the predictors, even if our target is to predict gene expressions at the cell-type level. As shown in Supplementary Fig. 9, increasing the number of cells for training also increased the gene-level correlations. We also compared the prediction results between the model input as pseudo-bulk information and the model input as single-cell information to further support our conclusions, shown in Supplementary Fig. 10a, b, for these two datasets. We investigated the selected sequence length's effect on prediction performances, summarized in Supplementary Note 1. Our model can also predict gene expression levels for atlas-level datasets, which is further discussed in Supplementary Note 2.

https://doi.org/10.1038/s41467-025-64506-8

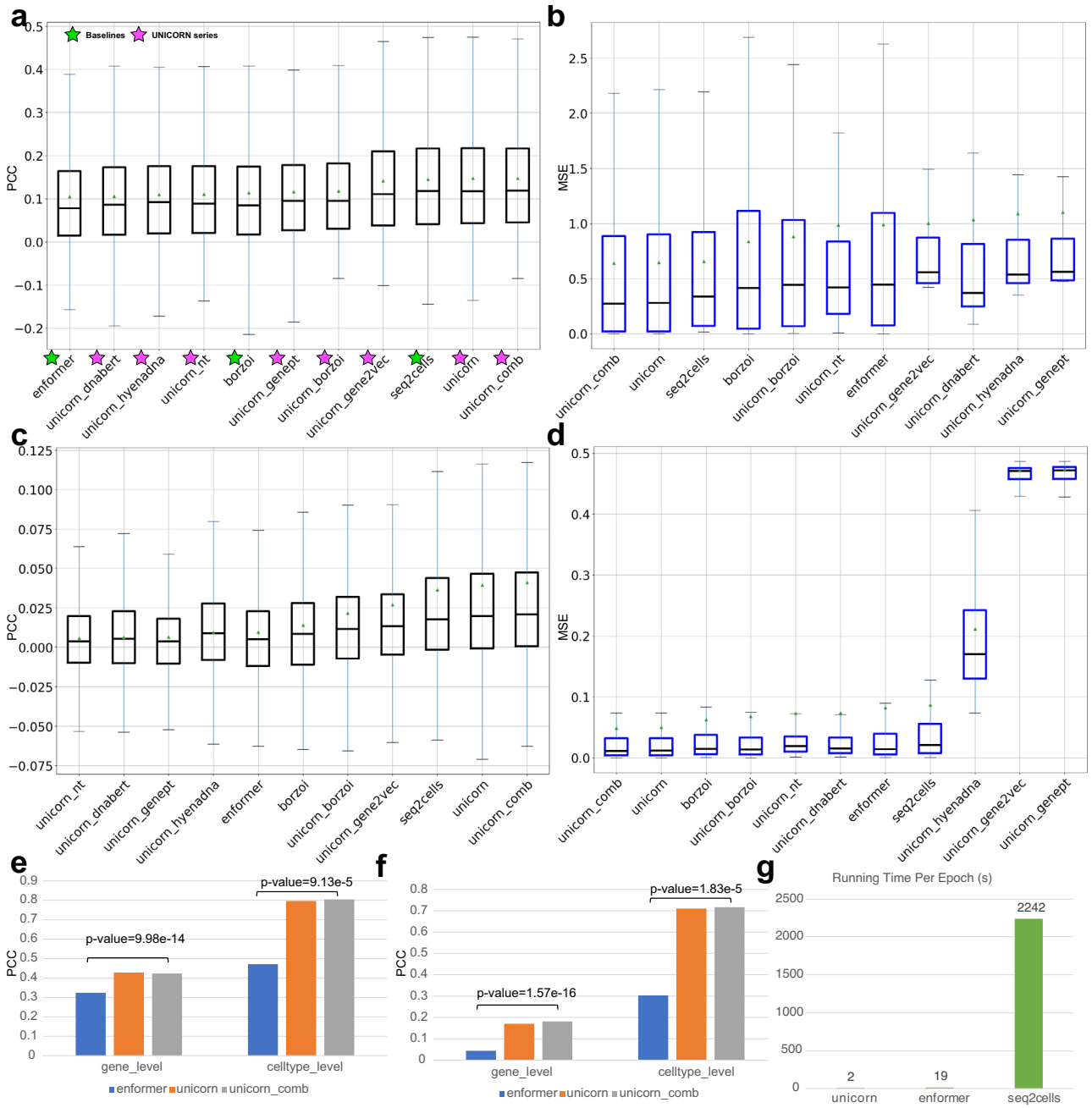


Fig. 2 | Comprehensive evaluations for gene expression predictions from sequences. The triangle shape represents the mean value, and the black dashed line represents the median value. We report the results based on box plots. **a** Results of gene-level correlations across different methods based on the thymus dataset ($n = 1766$ per group; triangle, mean; center line, median; box limits, upper and lower quartiles; whiskers, up to 1.5 × interquartile range; points, outliers). **b** Results of MSE across different methods based on the thymus dataset ($n = 1766$ per group; triangle, mean; center line, median; box limits, upper and lower quartiles; whiskers, up to 1.5 × interquartile range; points, outliers). **c** Results of gene-level correlations across different methods based on the PBMC dataset ($n = 1228$ per group; triangle, mean; center line, median; box limits, upper and lower quartiles; whiskers, up to 1.5 × interquartile range; points, outliers). **d** Results of MSE across different methods based on the PBMC dataset ($n = 1228$ per group; triangle, mean; center line, median; box limits, upper and lower quartiles; whiskers, up to 1.5 × interquartile range; points, outliers). **e** Results of gene-level correlations with pseudo-bulk evaluation based on the thymus dataset. The test is a two-sided Wilcoxon Rank-Sum test. **f** Results of gene-level correlations with pseudo-bulk evaluation based on the PBMC dataset. The test is a two-sided Wilcoxon Rank-Sum test. **g** Training time per epoch of different models. The unit is second (s). Source data are provided as a Source Data file.

Uncovering the complexity of model training

Explainability is an important part of understanding how neural networks work. In this section, we utilize function-based explainability to understand the factors that affect the performance of predictions using the model's output based on different scenarios[48]. Here, we discuss the first component of explainability, which helps us understand the behavior of gene expression prediction. In Fig. 3a with stages ①-③, we illustrate the differences in model performance under different settings. We can adopt many ideas to improve model performance, and the basic approach is to tune hyperparameters and test a better loss function design. Such a design can explain the performance improvement in ①. Secondly, we consider addressing the sparsity problem existing in single-cell datasets. In the past, seq2cells utilized scVI[49] to impute the scRNA-seq dataset as a denoising step.

https://doi.org/10.1038/s41467-025-64506-8

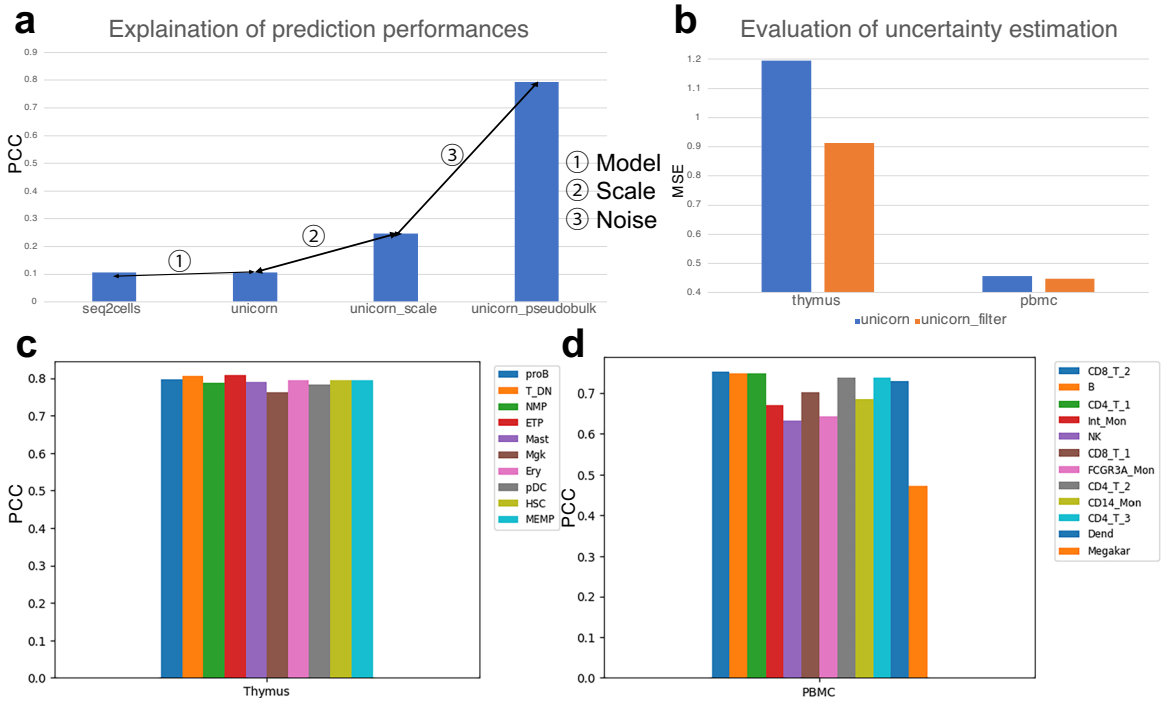


Fig. 3 | Explainability of UNICORN helps understand the principles of expression prediction. a Explainability of UNICORN from the differences of pre-processing steps and post-processing steps. **b** Explainability of UNICORN from uncertainty estimation. **c** Average correlation coefficients across different cell types for the thymus dataset. **d** Average correlation coefficients across different cell types for the PBMC dataset. Source data are provided as a Source Data file.

However, such a design is biased to the imputation method because we found that using different denoising methods led to different results, shown in Supplementary Fig. 11, including the comparison between scVI and DeepImpute[50]. Therefore, handling sparsity might not fully address this problem, so we standardized the single-cell dataset using z-norm for each gene to reduce the existence of zeros, as the neural network is hard to learn and predict zeros in the regression problem[51]. Such a design can explain the performance improvement in ②. Finally, the samples in the single-cell dataset can be treated as replicates of specific cell types, and predicting single-cell gene expression is still difficult since our previous improvement is insufficient. Therefore, we evaluate the outputs of UNICORN by generating the pseudo-bulk dataset aggregated by cell types and visualize the result in Fig. 3a. We found that this method led to noticeable improvement, and the performance gap in ③ can be explained by data complexity. Furthermore, we also tried to explore the linkage between gene functions and expression prediction performance, which are further discussed in Supplementary Note 3.

In conclusion, sequence-based models are promising for predicting cell-type-specific gene expression levels powered by methods from deep learning.

Improving prediction performance by uncertainty estimation

Here, we discuss the second component of explainability, which facilitates the selection of more informative genes inspired by the uncertainty estimation framework[52]. Our idea is to estimate the uncertainty levels of each sample to identify the reliability of the model prediction. We computed the uncertainty of each gene and each cell based on the loss prediction method (LossPred)[53]. As shown in Fig. 3b, by filtering the genes whose uncertainty scores were higher than the median (denoted as a high-uncertainty group), we achieved lower MSE for both the thymus and PBMC datasets. Furthermore, to validate the relationship between the uncertainty levels of each gene as well as its statistics, we computed the variance across cells for genes in the high-uncertainty group and low-uncertainty group, and the average values are shown in Supplementary Fig. 12. We found that genes with low uncertainty also had low variance on average, and thus our estimation of uncertainty is also supported by the features of tested datasets.

Moreover, we explored the possibility of explaining the uncertainty scores from biological perspectives. Here, we selected the remaining genes and ran Gene Ontology Enrichment Analysis (GOEA)[54,55], recorded in Supplementary Data 1. The adjusted *p*-value threshold of significance is set as 0.05. For both datasets, the gene sets with lower uncertainty always had more significantly enriched pathways (37 pathways from the low uncertainty group, 28 pathways from the high uncertainty group in PBMC; 148 pathways from the low uncertainty group, 112 pathways from the high uncertainty group in thymus.), which supports our findings that these genes play a more important role in biological activity. Moreover, genes with different uncertainty levels still have functional similarity reflected by the overlaps of pathways, including regulation of cellular metabolic process (GO:0031323), RNA metabolic process (GO:0016070), and others, while genes with lower uncertainty demonstrate more enrichment related to tissue-specific functions, supported by the tissue-specific pathway discovery shown in Supplementary Fig. 13. For example, the discovered pathways related to myeloid development represents one of the most important thymus functions. We also analyzed the relationship between the coefficient of variation (CV) based on expression profiles for two gene sets, shown in Supplementary Fig. 14a, b for different datasets. We found that genes with lower uncertainty also have higher CV, which implies that genes with larger CV values may carry biological significance. In conclusion, our uncertainty estimation component can help us identify difficult components of expression prediction.

Analyzing effects from cell types and individuals for expression prediction

Previous work suggests that the expressions and functions of the same gene vary across different cell types, which derived specific functions of cell types and tissues[56–58]. Therefore, we investigate if our model has different performances for different cell types. To conduct this experiment, we trained different models for different cell types to compute the MSE and correlation across genes and cells, and visualize our metrics in Fig. 3c for the thymus and in Fig. 3d for PBMC. For the thymus dataset, our model achieved the best performance in predicting gene expressions for the ETP cells, and the worst for the Mgk cells, with no significant differences across different cell types. However, for the PBMC dataset, our model performed best for the CD8 T 2 cells, and worst for the Megakar cells, with significant differences between these two cell types. To explore factors associated with prediction performance, we calculated the PCCs between the number of cells for different cell types and the corresponding MSE metric for both datasets, and they both showed significant negative correlations (corr = − 0.83, p-value = 0.003 for the thymus dataset; and corr = − 0.63, p-value = 0.03 for the PBMC dataset). Therefore, UNICORN performs better in predicting gene expressions for major cell types, and increasing the number of sequenced cells can potentially improve the performance of UNICORN.

Although UNICORN showed promising performance in predicting gene expression for certain cell types, it might be more reliable to replace the sequences from the reference genome with sequences measured directly from the individuals. Therefore, we investigated whether UNICORN can perform individual-level gene expression predictions based on paired whole-genome sequencing (WGS)[59] data for individuals. Here, we collected all the scRNA-seq datasets as well as corresponding WGS data from the Genotype-Tissue Expression project (GTEx) v9[60]. The database from GTEx v9 contains 16 individuals with their corresponding WGS and scRNA-seq datasets. First, we selected the gene expression profiles from the overlapped cell type (endothelial cell) across individuals and trained one model for each person. We also controlled the same training/validation/testing sets of genes across individuals and incorporated Enformer without fine-tuning as a baseline by selecting the output track of the endothelial cell as its output. The predicted values are evaluated based on the pseudo-bulk mode. In Supplementary Fig. 16, we showed the performances of UNICORN in predicting the unseen genes for each person. UNICORN showed more similar expression patterns as well as consistency across individuals compared with observed expression levels, while Enformer failed to reflect the personal expression levels, shown in Fig. 4a. Our findings also matched previous conclusions on the drawbacks of using Enformer for individualized gene expression prediction[61,62]. To strengthen our conclusion, we also fine-tuned Enformer based on the selected sample 12BJ1 (which is the sample with the highest averaged correlation predicted by UNICORN) and visualized the comparison between fine-tuned Enformer and UNICORN in Supplementary Fig. 15a. According to this figure, UNICORN also had lower MSE and higher cell-type correlation. Examples of genes with good (high correlation) and poor predictions (low correlation) are shown in Supplementary Fig. 17a, b. Therefore, the prediction of individualized gene expression at the cell-type level is practical, and UNICORN is a better model for learning the linkage between sequence and expression. Furthermore, we investigated the relationship between data quality and prediction performance to study the influence of technical noise on biological-related conclusions. We collected the cell type information of 16 different scRNA-seq datasets and computed the clustering scores by averaging Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI) to evaluate the quality of the given datasets, shown in Supplementary Fig. 15b. According to this figure, we identify four samples (15EOM, 1MCC2, 1CAMS, and 1CAMR) whose clustering scores are lower than other samples' scores. We then trained UNICORN based on the whole scRNA-seq dataset of each person and computed the correlations and MSE between observed and predicted gene expression profiles across individuals, shown in Supplementary Fig. 15c, d. According to these two figures, we identified that samples with low correlations on average were also those with low data quality. These four samples also have a low cell-type-level correlation of the endothelial cell, shown in Fig. 4a. After filtering these samples, we tested if the model trained by one sample could be used to predict the gene expression on endothelial cells of other samples, and the results are shown in Fig. 4b and Supplementary Fig. 17c. According to these figures, the performance of different models is stable for a given testing sample, which also shows the potential of UNICORN in learning the shared genomic features across individuals. Therefore, UNICORN's poor performances in certain samples are explained by their low data quality, and thus, collecting scRNA-seq data with good cell-type stratification and paired WGS information should enhance the performance of UNICORN in predicting gene expression across different dimensions.

Since our current model accepts individualized genome information as inputs and individualized gene expression levels as outputs, our model can also be used to evaluate the effects of specific variants[63]. Our discussion can be found in Supplementary Note 6. Although UNICORN is a good model for predicting gene expression levels, this method still cannot explain the variant effects well due to the group-level effects of variants, as well as the pre-training setting of Enformer.

Predicting multi-omic phenotypes at cell-type resolution simultaneously

The capacity of UNICORN allows us to jointly predict multi-omic phenotypes from one cell type, which could help identify regulatory relationships between different biological features. In this section, we utilized one 10X Multi-omic dataset containing gene expression and ATAC peak profiles, and one CITE-seq dataset containing gene expression and surface protein expression profiles for experiments. For the 10X Multi-omic dataset, the input of UNICORN came from the genomic sequence embeddings generated by Enformer. For the CITE-seq dataset, the input of UNICORN comes from the genomic sequence embeddings generated by Enformer and the amino acid sequence from ESM2. We hypothesized that including multi-omic information can help the prediction of each modality, relative to the single-modality case. Therefore, we compared the prediction performances with only one modality with joint modalities, shown in Fig. 5a for the 10x Multi-omic dataset and Fig. 5b for the CITE-seq dataset. According to these two figures, having more modalities can help predict different modalities, especially for predicting the surface protein expression levels across cell types. We also tested various settings of embeddings and did not find significant differences, shown in Supplementary Fig. 18a. Therefore, our current setting is capable of predicting the expression levels in the multi-omic setting.

We also hypothesized that UNICORN can offer insights into selecting important features for downstream analysis. Here, we selected the peak measured between 8417393 and 8418413 on chromosome 1, as its predicted value is highly correlated with the observed value (corr > 0.99). We visualized the relationship between the predicted and observed values for this peak in Supplementary Fig. 18b. Furthermore, this gene overlaps with NDEL1, as shown in Fig. 5c. NDEL1 expressed in monocytes has proved to be able to regulate osteoclasts and bone homeostasis[64], and thus UNICORN can accurately predict the important peak associated with biological functions.

Predicting disease effects and perturbation effects uncovers the biological variation

By performing predictions for gene expression levels under different conditions (including diseased/non-diseased and perturbed/non-perturbed), UNICORN can also help us understand the biological

https://doi.org/10.1038/s41467-025-64506-8

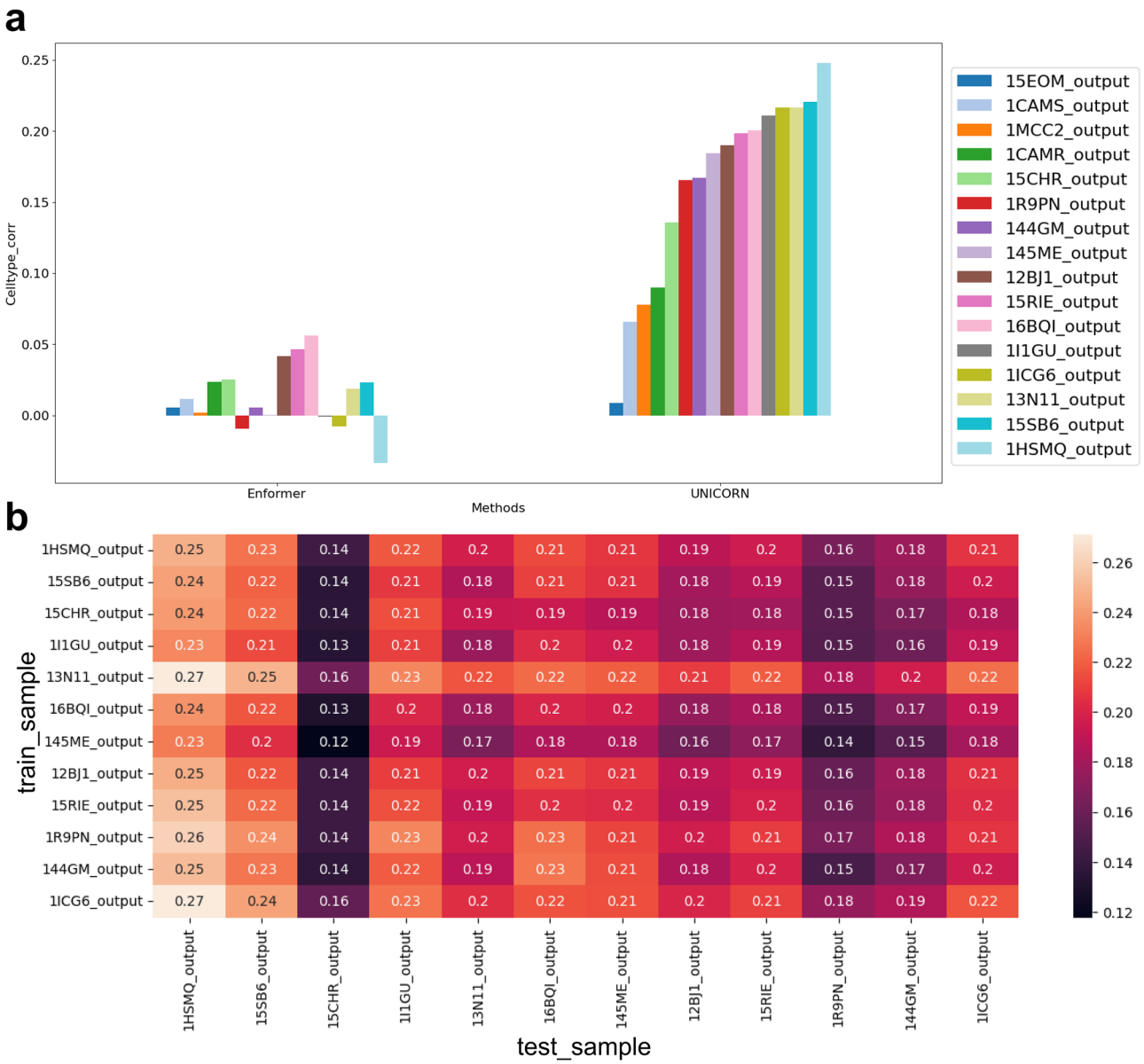


Fig. 4 | Results of gene expression predictions for individualized scRNA-seq datasets. The shared cell type is an endothelial cell. **a** The results of cell-type correlations computed based on the given cell type across individuals for outputs from Enformer and UNICORN. Samples are ranked based on the cell-type correlation from UNICORN in ascending order. **b** The cell-type correlation of cross-individual prediction results based on UNICORN, after filtering the low-quality samples. The rows represent training samples and the columns represent the testing samples. Source data are provided as a Source Data file.

differences of cells and genes with respect to observed conditions. Following our previous design for analyzing the differences in prediction results across different cell types, we predicted the expression levels of testing genes based on three datasets from different conditions. The datasets Aorta[65] and Heart[66] are used to investigate the effects of diseases, and the dataset Openproblem[67] is used to investigate the effects of perturbations. The Aorta dataset has four disease states, including Ascending only, Ascending to descending, Ascending w/root, and Control. The Heart dataset has three disease states, including hypertrophic (HCM), dilated cardiomyopathy (DCM), and non-failing (NF). We utilized the default hyperparameters to fine-tune different models for these datasets and compared the performances for the prediction results conditioned by disease states and perturbations. Our prediction results at the pseudo-bulk level are aggregated based on conditions and evaluated based on both cell-level and gene-level metrics.

First, we analyzed the differences in prediction results across different diseased and non-disease conditions. Figure 6a shows the prediction results based on the Aorta dataset, and the labeled results, such as “Subset: Control" represents that our trained model is only based on cells under the condition Control. According to this figure, UNICORN had the worst performance in predicting cells under the ascending to descending condition, which is an intermediate state between the rest of the states. Such a finding implies that cells in the intermediate state might be hard to predict because cells under this condition are not stable. Figure 6b shows the prediction results based on the Heart dataset, and the labeled results like “Subset: NF" represent that our trained model is only based on cells under the condition NF. According to this figure, UNICORN had the worst performance in predicting cells under the NF condition, which is unexpected. Such findings might be explained by the differences in data quality. Meanwhile, we assume that incorporating cells with different phenotypes

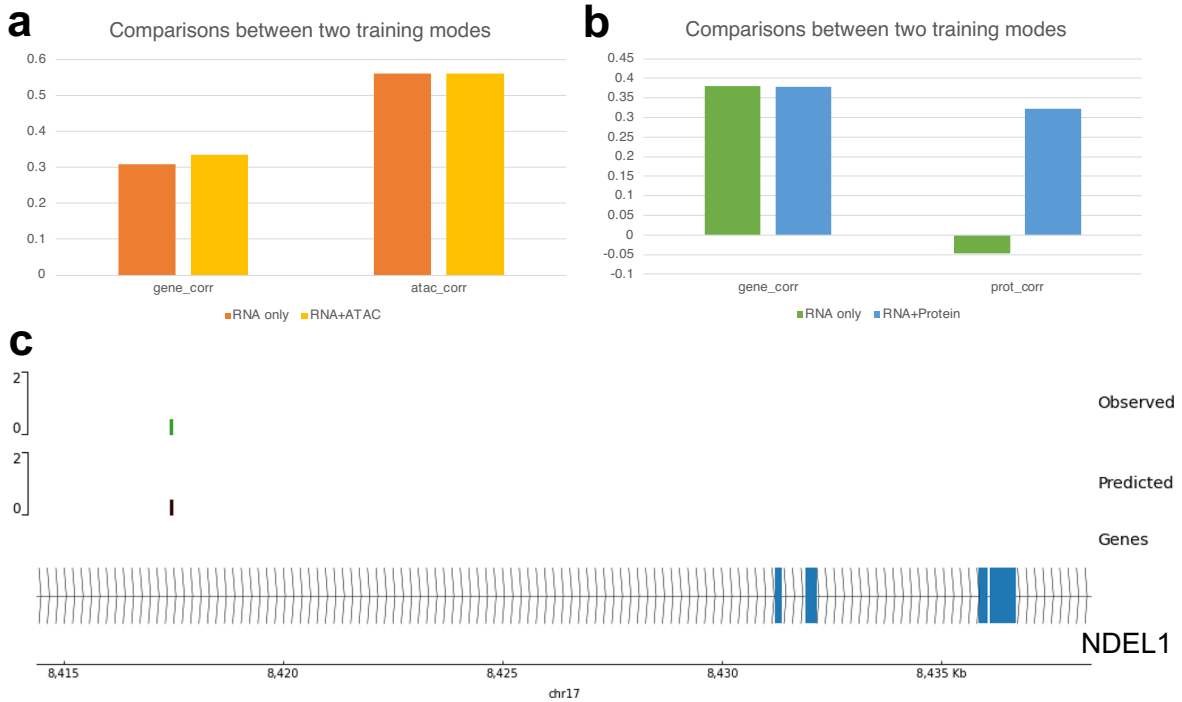


Fig. 5 | UNICORN demonstrates its ability in predicting multi-omic information at the cell-type level. Here, RNA represents the gene expression profiles. The evaluation is performed based on the pseudo-bulk data. We reported the averaged correlation scores in (**a**) and (**b**). **a** The comparisons between RNA-only mode or ATAC only mode with the joint training mode (RNA + ATAC). **b** The comparisons between RNA-only mode or Protein-only mode with the joint training mode (RNA + Protein). **c** Visualization of observed and predicted peak information for CD 14 Mono cells. We selected the peak with the highest correlation for analysis. Source data are provided as a Source Data file.

can also enhance the overall prediction performance, and to validate if training with diseased cells can facilitate the prediction for normal cells or not, we include the prediction results labeled with “Subset: Control" and “Subset: NF" in our discussion. We found that for both datasets, including diseased cells can help predict gene expressions from normal cells and vice versa. Therefore, including heterogeneity in the training process can improve the performance of UNICORN. There is a similar conclusion from ref. 68 related to cell-type transferring. Furthermore, UNICORN can preserve the expression levels of most differential expression genes (DEGs) under different disease states based on the testing split, validated by the strong correlation. For example, we observed accurate prediction results for genes RPL27A (corr = 0.96), RPL28 (corr = 0.76), and SON (corr = 0.96) from the Aorta dataset, and gene NT5C2 (corr=0.99) for the Heart dataset. Notably, the variants in NT5C2 have been shown to have a high risk in the association with heart diseases, and thus UNICORN also shows strength in capturing the correct expression levels for these disease-associated genes[69]. Therefore, UNICORN can help us analyze the differences of various conditions and suggest the importance of collecting cells from different states to build atlas-level databases.

We then investigated the performance of UNICORN in predicting gene expression levels under different perturbations. Figure 6c shows the gene-level correlation results under three different cases, known as Ganetespib (STA-9090) (a perturbation with the highest corr score), Ketoconazole (a perturbation with the lowest corr score), and Average (the average corr score across all the perturbations). Such a difference is also reflected in the results of the predicted DEG values under the perturbation cases, as the DEG SRSF7 of Ganetespib (STA-9090) had a high correlation score (gene_corr = 0.70) and the DEG MX1 of Ketoconazole had a low correlation score (gene_corr = -0.77). Figure 6d shows the MSE results under three different cases, known as CGP 60474 (a perturbation with the lowest MSE score), Oprozomib (ONX 0912) (a perturbation with the highest MSE score), and Average (the averaged MSE score across all the perturbations). The predicted value of DEG SRSF7 from CGP 60474 had a high correlation (gene_corr = 0.99), and DEG SNHG6 from Oprozomib (ONX 0912) also had a low correlation score (gene_corr = − 0.83). The ranks of the same perturbations under different metrics can also be different because of the existing intrinsic biases in evaluations. The value of metrics represents the difficulties of predicting these perturbations’ effects on gene expression levels based on the testing split, which can also be used to describe the heterogeneity of these perturbations. Therefore, the result of UNICORN can also offer a new direction to model the perturbation effects from a computational perspective.

Perturbation or disease effect can lead to the change of gene expression levels in both direction and magnitude. Therefore, it is also important to test if UNICORN can capture such a difference in the predicted expression profiles across different conditions. We have provided a comparison of condition-level gene expression differences between predicted and observed levels. According to Supplementary Fig. 19a, b, the difference computed based on predicted and observed data is similar and has a significant correlation for selected disease datasets. For the perturbation dataset, since we have collected over 100 perturbations, we display the correlation between the difference computed based on the predicted and observed levels of each perturbation in Supplementary Fig.19c. This figure also shows that all the correlations are positive, and the averaged PCC is also high (> 0.3). Therefore, we can conclude that our model can correctly reflect differences in conditional gene expression profiles.

Discussion

Predicting gene expression from DNA sequence directly is a long-standing and challenging problem in biology. An effective model can help researchers better understand the relationship between

https://doi.org/10.1038/s41467-025-64506-8

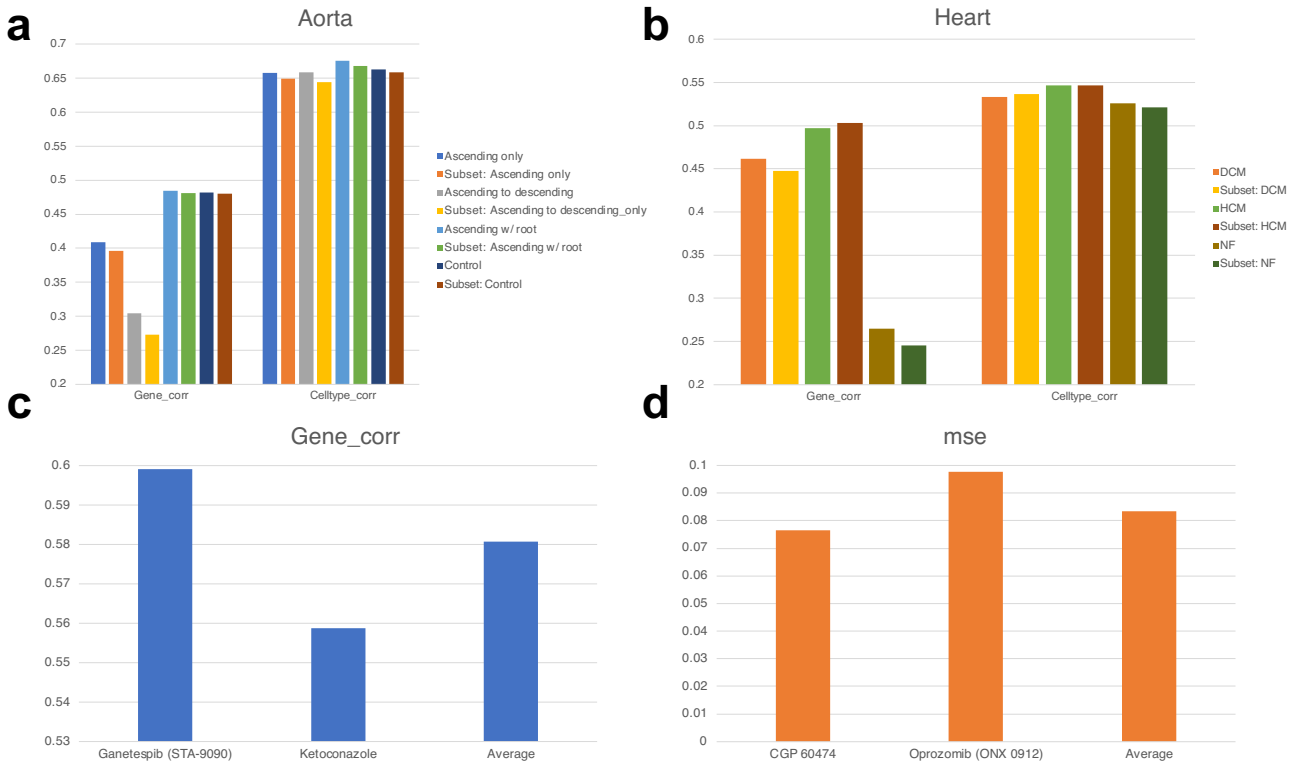


Fig. 6 | UNICORN uncovers the complexity of diseases and perturbations. a The correlation coefficients computed based on genes and cells across different conditions from the Aorta dataset. **b** The correlation coefficients computed based on genes and cells across different conditions from the Heart dataset. **c** Visualization of selected perturbation conditions by correlation coefficients computed based on genes. We only include perturbations with the highest and lowest correlation coefficients, as well as the averaged value across all perturbations. **d** Visualization of selected perturbation conditions by MSE computed based on genes. We only include perturbations with the highest and lowest MSE, as well as the averaged value across all perturbations. Source data are provided as a Source Data file.

regulatory elements and gene expressions at the cell-type or cell-state levels. In this work, we present UNICORN, to address this research question with the help of prior biological information as well as an advanced machine learning framework. UNICORN utilizes pre-trained sequence-to-function models based on non-single-cell-level data as well as LLMs and is fine-tuned based on single-cell multi-omic datasets to perform predictions under different resolutions. Furthermore, UNICORN is capable of explaining the prediction process from different aspects. UNICORN not only can predict the multi-omic information but can also give individualized outputs based on DNA sequences from cohort studies and help in analyzing the perturbation effects based on diseased or perturbed data. Overall, UNICORN can improve prediction performances as well as gain biological insights on expression regulation.

Regarding the contributions of UNICORN for single-cell multi-omic data prediction, we demonstrated the feasibility of prediction of molecular expressions at the single-cell and cell-type resolutions, but the performance depends on data quality and requires attention to avoid overfitting. Although we demonstrated UNICORN's better performances by running a comprehensive benchmarking analysis, it will always be helpful to explore pre-trained sequence-based models and possibly use them to enhance prediction performance. It is also interesting to note that pre-trained models focusing on gene expression prediction in other types of data performed better than other foundation models trained directly with DNA sequences in this task. Meanwhile, as expected, predicting gene expression at the cell-type level is easier than that at the single-cell level, especially for major cell types, which are further validated in our experiments for reducing the sparsity of datasets and increasing the number of cells used for training. By estimating the uncertainty levels of predicted genes, we can further offer more reliable prediction results. Therefore, UNICORN can help us uncover the heterogeneity of (gene) expression prediction at different resolutions.

Previous sequence-to-expression models utilize the reference genome as input data and thus unpaired expression profiles are used for training. Here, we investigated the capacity of UNICORN in predicting individualized gene expression profiles with the help of paired WGS data from the GTEx cohort. The paired data also incorporate the variant information of each individual, and thus, it is a better resource than those models trained with unpaired information. Based on our analysis, the performance of UNICORN varies from person to person. Moreover, the pre-trained model (Enformer) used in this section is still from the reference genome, and we lack such a model pre-trained with individual information and thus, it will be helpful to inherit the structure of UNICORN with an individual-aware sequence-based model for this task.

Inspired by our observations of the heterogeneity of performance across different cell types, we consider utilizing UNICORN to uncover disease effects and perturbation effects by analyzing the predicted gene expression across different conditions. Based on our experiments, including cells from more disease states can enhance the prediction performances under different conditions, even for the control case. Therefore, UNICORN can also be used to demonstrate the necessity of sequencing and collecting large-scale single-cell datasets under different disease states, which are valuable resources for model improvement. Regarding the analysis of perturbation effects, UNICORN showed different performances for different genes, which could be further used to investigate the strength of perturbation effects on cellular activity. Therefore, UNICORN can be used to explore the unique signals existing in different perturbations.

Despite the promising results from UNICORN, there is still a big gap between the DNA sequence and the corresponding expression levels. Firstly, the stochastic nature of gene expression regulation at the single-cell level, coupled with high noise in expression measurements pose a major challenge to accurately construct the cell-type-specific gene expression, so there is a need to collect more high-quality cells to improve prediction accuracy. Secondly, there are various approaches to pre-train a base model from sequence information for this task, and we are still looking for the best framework which can capture both the context of sequence information, the gene functional information, as well as the regulators for gene expression changes. Considering the inclusion of cohort-level data for enriching the corpus of the training set may be an approach. Moreover, UNICORN does not accurately capture the relative changes for certain diseases, which implies the necessity of dataset-specific fine-tuning with the base model. Finally, unraveling the contributions of variant effects at the single-cell or cell-type levels is still an important and challenging task, and the current framework may not fundamentally solve it due to the limitation of design. For example, we did not have models pre-trained with gene expression levels from individuals for different cell states, and we might need a better model architecture other than transformer or an improved loss function to capture the contribution of variants in controlling the expression levels or other related phenotypes. Therefore, exploring the contribution of a pre-training + fine-tuning framework for explaining such effects is still welcoming for further exploration.

Overall, UNICORN is a useful tool for unifying the expression prediction tasks from the sequence-based information and understanding the complex biology system created by the effects of genetic variation.

Methods

Problem statement

Considering we have a single-cell dataset $\mathcal{D}$ with modalities $1 \ldots m$, we then can denote the single-cell dataset as $\mathcal{D} = \{X_1^{n \times p_1}, X_2^{n \times p_2}, \ldots, X_m^{n \times p_m}\}$. Each expression matrix X has the same number of sequenced cells n but a different number of biological features $p_1 \ldots p_m$. These expression matrices represent our prediction targets. Considering we also have biological feature embeddings for all modalities from pre-training models, which are denoted as $\mathcal{P} = \{e_1^{d_1 \times p_1}, e_2^{d_2 \times p_2}, \ldots, e_2^{d_m \times p_m}\}$. Our target is to fine-tune two models f_{θ_1} to predict the targeted expression levels and g_{θ_2} to estimate the uncertainty of prediction, where $\theta_1$andθ_2 represent two sets of parameters, that is:

$$\begin{aligned} \hat{X}_i^{n \times p_i} &= f_{\theta_1}(e_i^{d_i \times p_i}), \\ \hat{U}_i^{n \times p_i} &= g_{\theta_2}(e_i^{d_i \times p_i}), \end{aligned} \tag{1}$$

where $\hat{X}$ represents the predicted gene expression and $\hat{U}$ represents the estimated uncertainty. Our equation utilizes the i^{th} modality as an example. The output dimension of f_{θ_1} is equal to the number of cells, as we are predicting gene expression levels from a sequence given the context of a specific dataset.

Explanations of model architecture

There are two components of UNICORN. The first component is a pre-trained model to extract the embeddings from biological features as inputs. Using gene as an example, we utilize Enformer to generate DNA embeddings based on DNA sequences of the given genes and the GPT 3.5 embedding layer of the text description of these genes from NCBI to generate text embeddings. The source species of the target datasets determines the genome of DNA sequences. We then combine DNA embeddings and text embeddings to finalize our dataset. We note that the generator of these two types of embeddings can be changed, so it should be updated with current state-of-the-art (SOTA) methods for each subproblem. Moreover, we use the same model to generate the embeddings of peaks for scATAC-seq data from DNA sequences and use ESM2[70] to generate the embeddings of proteins for CITE-seq data from amino acid sequences. ESM2 was demonstrated as a top performer in protein-related tasks[71]. After having the group of embeddings as input, we assign separate predictors for each modality, and each predictor is constructed based on a two-layer MLP. We utilize ReLU[35] as the activation function and Softplus[35] as the output layer. The former design ensures that our model can learn non-linear relationships, and the latter design ensures that our model always has positive prediction results.

Meanwhile, for each predictor, we have the corresponding uncertainty estimator g_{θ_2}. Our estimator and predictor have the same architectures but do not share weights because they work for different predictions. The method we used to estimate the uncertainty for prediction output is LossPred, which means we directly predict the MSE of the given example during the training process. This uncertainty estimation component will combine with the default loss function when we are interested in exploring the uncertainty quantification of different features. Such a method, supported by[52], outperforms other methods in their evaluation results across various datasets. Our current design is more efficient than full-parameter fine-tuning based on Enformer or other large-scale prediction models for genomic data, and we also demonstrate the superiority of our current design in the Results section.

Model explainability

We utilize function-based explainability to understand the factors that affect the performance of predictions and select critical genes. The idea of function-based explainability is designed to understand the contribution of sample-level information for making decisions based on deep-learning models, which has been widely discussed in several method-driven papers or systemic review papers[72–74]. This idea is also critical in biology research for us to understand hidden mechanisms in a transparent approach[75]. In our experiment, each sample is one gene so our target is to understand the role of gene-level signals in making the prediction. To achieve this target, we have designed several experiments for exploration. First, we consider processing the training data with different pre-processing methods or aggregating methods to investigate the progress of making more accurate predictions. This approach can be interpreted as an approach to explain the effect of distribution-level assumption for gene expression measurement. Second, we perform various data ablation studies to investigate the introduction of sparsity and cell-type proportion for model prediction. This approach can be interpreted as an approach to explain the effect of the modeling assumption for gene expression measurement. Third, we utilize the estimated uncertainty of biological features to select genes with low uncertainty and conduct biological analysis, as well as training dynamics analysis. Our uncertainty estimation approach has a comprehensive theoretical foundation supported by experiments with simulation and real transcriptomic data. This approach allows us to provide a computational-level explanation for each gene and uncovers the biological signals from different pathway-associated genes selected by UNICORN.

Details of model training

We first discuss our choices of loss functions. We utilize different loss functions for single-modality prediction and multi-modality prediction. For the single-modality prediction, we combine mutual Pearson correlation (MPC), Poisson negative log-likelihood loss (PNLL), and mean squared error (MSE) losses together to finalize our loss functions. Considering we have prediction $\hat{Y}_k$ and observed values Y_k for

modality k, our loss function is defined as:

$$\begin{aligned}
\mathcal{L}_{MPC} &= -\left(\frac{1}{n}\sum_{i=1}^{n} Pearson(Y_{k(i,:)}, \hat{Y_{k(i,:)}}) + \frac{1}{p_k}\sum_{j=1}^{p_k} Pearson(Y_{k(:,j)}, \hat{Y_{k(:,j)}})\right),\\
\mathcal{L}_{PNLL} &= \hat{Y_k} - Y_k \log(\hat{Y_k}),\\
\mathcal{L}_{MSE} &= \frac{1}{np_k}\sum_{i}^{n}\sum_{j}^{p_k}(Y_{k(i,j)} - \hat{Y_{k(i,j)}})^2,\\
\mathcal{L}_{total} &= \lambda_{MPC}\mathcal{L}_{MPC} + \lambda_{PNLL}\mathcal{L}_{PNLL} + \lambda_{MSE}\mathcal{L}_{MSE}.
\end{aligned} \tag{2}$$

Here, the function $Pearson(a, b)$ computes the Pearson correlation coefficients between two items, and our method is based on cosine similarity, and thus, it is differentiable. the notation (i, j) represents the i^{th} cell and j^{th} feature in the predicted matrix. The weights of these loss components, denoted by λ_X, are the same after tuning with different options. To elaborate on the necessity of including all three loss components to formalize our final loss function, we perform ablation tests shown in Supplementary Fig. 20a–c under different metrics.

For the foundation support of our loss function design, we referred from the theory of empirical risk[76] and multi-task learning[77] to provide further support. Our loss function considers three different tasks, and thus it covers multiple objectives when performing optimization. To predict high-dimensional biological data, relative patterns are more robust and biologically meaningful. Therefore, maximizing the correlation between predicted values and observed values can better align model predictions with the underlying biological structure. Furthermore, Pearson correlation is scale-invariant, so it can also help our method learn invariant representations that preserve structure rather than magnitude. Based on the observation of data sparsity and distribution, we introduce Poisson Negative Log-Likelihood loss (PNLL) for modeling data sparsity explicitly and model gene expression profiles. Finally, we include mean squared error loss to capture the magnitude-level similarity between predicted value and observed value, which further captures the value-level fidelity globally. Therefore, in our design, different components act as regularizers, guiding the model to learn representations that are simultaneously expressive, interpretable, and robust to noise and data scale.

For multi-modal prediction, instead of applying separate prediction heads for each modality, we predict the multi-modal information jointly for a single cell or a single cell type. We design the loss functions for multi-modal prediction based on a distribution-aware strategy. Since scATAC-seq is very sparse, we use the same loss function designed for regressing gene expression profiles. Moreover, after normalization, the proteomic part of CITE-seq data is dense, so we only keep the PCC loss. This principle of loss design is also supported by our ablation studies for multi-omic expression prediction. To elaborate on the necessity of including distribution-aware loss function design to formalize our final loss function, we perform ablation tests shown in Supplementary Fig. 20d–f under different metrics. Here we found that including all three components is helpful for predicting the combination of scRNA-seq and scATAC-seq, while the prediction for CITE-seq data cannot be benefited by loss function modification. This observation aligns well with our discussion of the distribution-aware loss function design strategy. We also demonstrate the contributions of multi-modal prediction by comparing it with prediction based on a single modality.

We also discuss the hyperparameter selection for UNICORN. Most of the hyper-parameters are referred from seq2cells and tuned based on model's performance computed on the validation dataset. Details of our discussion can be found in Supplementary Note 4.

Regarding the motivation and foundation of uncertainty estimation based on LossPred, we have provided theoretical support as well as validation based on simulation data in Supplementary Note 7.

Methods for performing biological analysis

To map the biological features (including genes and peaks) with their relevant sequence window, we utilize the construction function from ref. 78 to retrieve the location of such features in the chromosome, as well as the starting site and the ending site.

To retrieve the functions of genes, we downloaded the gene functional information of labels from ref. 79. These genes are classified into protein-encoding genes, non-coding RNA, etc. Details of the gene functional list can be found in Supplementary Fig. 24a.

To infer the functions of gene sets we are interested in, we utilize Gene Ontology Enrichment Analysis (GOEA)[54,55] to analyze genes of different sets. GOEA is based on Fisher's exact test to compare the difference between the selected gene sets and all background genes, with outputs including significance level and overlap ratios.

Information of metrics

To evaluate different models for expression predictions, we consider two metrics for two dimensions. The metrics include Pearson correlation coefficients and mean squared errors. These metrics are widely used in evaluating tools for gene expression prediction[11,21,24,27,80–82]. Since the sequencing data do not contain autocorrelated patterns, correlation can capture the similarity of tendency between predicted and observed expression levels, while MSE can capture the absolute similarity between predicted and observed expression levels. Therefore, the combination of metrics we choose provides a comprehensive measure of the difference between predicted and observed profiles. Our metrics are computed using the scikit-learn package[83] and based on testing genes. Our dimensions include gene-level evaluation and cell-level evaluation. The former setting represents that we compute the correlation coefficients across genes for each cell/cell type/phenotype, and the latter setting represents that we compute the correlation coefficients across cells/cell types/phenotypes for each gene. We report the mean, median, and quartiles as statistics shown in boxplots. The value of correlation is in $(-1, 1)$, and a higher value represents better model performance. The value of MSE is in $(0, \infty)$, and a lower value represents better model performance.

To evaluate different gene embeddings as well as cells for clustering, we consider Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and Averaged silhouette Width as metrics. All of them are in [0, 1], and higher values represent better performances.

Information of baselines

We consider different baselines for different tasks.

For the task of gene expression prediction, we include full-feature-based methods, including seq2cells[11], Enformer[21], Borzoi[24], Decima[22], GET[43], and variations of UNICORN as baselines. seq2cells is a model based on the sequence embeddings from pre-trained Enformer to predict gene expression levels at the single-cell resolution. Enformer and Borzoi are designed to predict gene expression from the corresponding sequence information, which allows us to fine-tune the full model to predict gene expression levels in single-cell resolution. Decima is pre-trained with the pseudo-bulk gene expression of large-scale single-cell transcriptomic data, and it can predict the gene expression of specific cell types under specific tissues. GET is pre-trained with (sc)ATAC-seq and it can predict gene expressions by fine-tuning this model with cell-type-specific information from scATAC-seq and scRNA-seq data. Our pipeline for fine-tuning Enformer and Borzoi follows the default design of gReLU. The variations of UNICORN come from different types of inputs. We also consider using embeddings from text descriptions of gene functions (genept)[38], embeddings from the DNA foundation model (HyenaDNA[2], DNABERT[29], NT[36]), and embeddings from gene-gene interaction (gene2vec)[37], as inputs. We also consider feature-selection-based methods, including SCOVER[40], DanQ[41], and LassoNet[42]. SCOVER is a convolutional neural network

(CNN) based method with motif-scoring design for expression prediction. DanQ integrates both CNN and a recurrent neural network (RNN) to model the functional information of genomic regions. LassoNet is a feature-selection-based method by penalizing the sparsity of features during the training process.

For the task of model explainability, we consider scVI[49] and DeepImpute[50] as baselines for imputing single-cell datasets. scVI is a model based on a variational auto-encoder to learn the cell embeddings and the distribution of gene expression levels in single-cell data. DeepImpute is a deep-learning-based method to impute missing gene expression in single-cell data by denoising auto-encoder.

For the task of multi-omic data prediction, we consider UNICORN with a shared predictor as the baseline model.

Data preprocessing

For the gene expression prediction task, we performed the log-normalized approach for scRNA-seq data based on scanpy[84]. We compared the performances of UNICORN between the raw mode and the normalized mode in Supplementary Note 5. For the multi-omic data prediction task, we used the raw count data from scATAC-seq and scaled gene expression data from scRNA-seq. Such a setting is relevant to our design of loss functions.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

We do not generate new data in this research, and all data used in this manuscript are publicly available without restricted access. The thymus dataset is available under accession code [https://zenodo.org/records/8314644]. The PBMC dataset is available under the access code [https://zenodo.org/records/8314644]. The thymus atlas dataset is available under the access code [https://zenodo.org/records/8314644]. The Aorta dataset is available under the access code [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE155468]. The Heart dataset is available under access code [https://singlecell.broadinstitute.org/single_cell/study/SCP1303/single-nuclei-profiling-of-human-dilated-and-hypertrophic-cardiomyopathy]. The Open-problems dataset is available under access code [https://www.kaggle.com/competitions/open-problems-single-cell-perturbations/data]. The Onek1k dataset is available under the access code [https://onek1k.org/]. For the datasets from GTEx, users need to submit an application for access under the access code [https://www.gtexportal.org/home/]. For the datasets from ROSMAP, users need to submit an application for access under the access code [https://adknowledgeportal.synapse.org/Explore/Studies/DetailsPage/StudyDetails?Study=syn3219045]. We include the links to access our experimental datasets in Supplementary Data 2. Source data are provided in this paper.

Code availability

To run UNICORN, we relied on Yale High-performance Computing Center (YCRC) and utilized one NVIDIA A40 GPU with up to 40 GB RAM. The upper bound of running time for the analysis is 24 h. The codes of UNICORN can be found in https://github.com/HelloWorldLTY/UNICORN and https://doi.org/10.5281/zenodo.16416845[85]. The license is MIT license.

References

1. Yang, Z. A space-time process model for the evolution of DNA sequences. *Genetics* **139**, 993–1005 (1995).
2. Nguyen, E. et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems* **36** (2024).
3. Reinagel, A. & Bray Speth, E. Beyond the central dogma: model-based learning of how genes determine phenotypes. *CBE Life Sci. Edu.* **15**, ar4 (2016).
4. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods* **12**, 931–934 (2015).
5. Zhou, J. et al. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* **50**, 1171–1179 (2018).
6. Ramprasad, P., Pai, N. & Pan, W. Enhancing personalized gene expression prediction from dna sequences using genomic foundation models. *Hum. Genet. Genom. Adv.* **5**, https://doi.org/10.1016/j.xhgg.2024.100347 (2024).
7. Drusinsky, S., Whalen, S. & Pollard, K. S. Deep-learning prediction of gene expression from personal genomes. Preprint at https://doi.org/10.1101/2024.07.27.605449 (2024).
8. Beer, M. A. & Tavazoie, S. Predicting gene expression from sequence. *Cell* **117**, 185–198 (2004).
9. Clark, W. T. & Radivojac, P. Analysis of protein function and its prediction from amino acid sequence. *Proteins* **79**, 2086–2096 (2011).
10. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature* 630, 493–500 (2024).
11. Schwessinger, R., Deasy, J., Woodruff, R. T., Young, S. & Branson, K. M. Single-cell gene expression prediction from DNA sequence at large contexts. *Preprint at* https://doi.org/10.1101/2023.07.26.550634 (2023).
12. Han, X. et al. Construction of a human cell landscape at single-cell level. *Nature* **581**, 303–309 (2020).
13. Saliba, A.-E., Westermann, A. J., Gorski, S. A. & Vogel, J. Single-cell rna-seq: advances and future challenges. *Nucleic Acids Res.* **42**, 8845–8860 (2014).
14. Lance, C. et al. Multimodal single cell data integration challenge: Results and lessons learned. In *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*,176 of *Proceedings of Machine Learning Research*, 162–176 (PMLR, 2022).
15. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
16. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
17. Brennecke, P. et al. Accounting for technical noise in single-cell rna-seq experiments. *Nat. Methods* **10**, 1093–1095 (2013).
18. You, Y. et al. Modeling group heteroscedasticity in single-cell rna-seq pseudo-bulk data. *Genome Biol.* **24**, 107 (2023).
19. Nathan, A. et al. Single-cell eqtl models reveal dynamic t cell state dependence of disease loci. *Nature* **606**, 120–128 (2022).
20. Liu, T. et al. Cosgenegate selects multi-functional and credible biomarkers for single-cell analysis. *Brief. Bioinform.* **26**, bbae626 (2025).
21. Avsec, Ž. et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
22. Lal, A. et al. Decoding sequence determinants of gene expression in diverse cellular and disease states. Preprint at https://doi.org/10.1101/2024.10.09.617507 (2024).
23. Kelley, D. R. et al. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.* **28**, 739–750 (2018).
24. Linder, J., Srivastava, D., Yuan, H., Agarwal, V. & Kelley, D. R. Predicting rna-seq coverage from dna sequence as a unifying model of gene regulation. *Nat. Genet.* **57**, 949–961 (2025).
25. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell rna-sequencing for biomedical research and clinical applications. *Genome Med.* **9**, 1–12 (2017).
26. Li, J. et al. Deep learning of cross-species single-cell landscapes identifies conserved regulatory programs underlying cell types. *Nat. Genet.* **54**, 1711–1720 (2022).

27. Hingerl, J. C. et al. scooby: Modeling multi-modal genomic profiles from dna sequence at single-cell resolution. Preprint at https://doi.org/10.1101/2024.09.19.613754 (2024).
28. Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D. & Makedon, F. A survey on contrastive self-supervised learning. *Technologies* **9**, 2 (2020).
29. Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics* **37**, 2112–2120 (2021).
30. Nguyen, E. et al. Sequence modeling and design from molecular to genome scale with evo. *Science* **386**, eado9336 (2024).
31. Marin, F. I. et al. Bend: Benchmarking dna language models on biologically meaningful tasks. In *The Twelfth International Conference on Learning Representations* (2023).
32. Bepler, T. & Berger, B. Learning the protein language: Evolution, structure, and function. *Cell Syst.* **12**, 654–669 (2021).
33. Zhao, W. X. et al. A survey of large language models. Preprint at https://doi.org/10.48550/arXiv.2303.18223 (2023).
34. Nachtomy, O., Shavit, A. & Yakhini, Z. Gene expression and the concept of the phenotype. *Stud. Hist. Philos. Biol. Biomed. Sci.* **38**, 238–254 (2007).
35. Paszke, A. et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).
36. Dalla-Torre, H. et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nat. Methods* **22**, 287–297 (2024).
37. Du, J. et al. Gene2vec: distributed representation of genes based on co-expression. *BMC Genom.* **20**, 7–15 (2019).
38. Chen, Y. & Zou, J. Simple and effective embedding model for single-cell biology built from ChatGPT. *Nat. Biomed. Eng.* **9**, 483–493 (2024).
39. Kang, B. et al. Decoupling representation and classifier for long-tailed recognition. In *The Eighth International Conference on Learning Representations* (2020).
40. Hepkema, J. et al. Predicting the impact of sequence motifs on gene regulation using single-cell data. *Genome Biol.* **24**, 189 (2023).
41. Quang, D. & Xie, X. Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences. *Nucleic Acids Res.* **44**, e107–e107 (2016).
42. Lemhadri, I., Ruan, F., Abraham, L. & Tibshirani, R. Lassonet: A neural network with feature sparsity. *J. Mach. Learn. Res.* **22**, 1–29 (2021).
43. Fu, X. et al. A foundation model of transcription across human cell types. *Nature* **637**, 965–973 (2025).
44. Yazar, S. et al. Single-cell eqtl mapping identifies cell type–specific genetic control of autoimmune disease. *Science* **376**, eabf3041 (2022).
45. Bennett, D. A. et al. Religious orders study and rush memory and aging project. *J. Alzheimer's Dis.* **64**, S161–S189 (2018).
46. Fang, R. et al. Comprehensive analysis of single cell atac-seq data with snapatac. *Nature Commun.* **12**, 1337 (2021).
47. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with signac. *Nat. Methods* **18**, 1333–1341 (2021).
48. Saranya, A. & Subhashini, R. A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. *Decision Anal. J.* **7**, 100230 (2023).
49. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
50. Arisdakessian, C., Poirion, O., Yunits, B., Zhu, X. & Garmire, L. X. Deepimpute: an accurate, fast, and scalable deep neural network method to impute single-cell rna-seq data. *Genome Biol.* **20**, 1–14 (2019).
51. De Cnudde, S., Ramon, Y., Martens, D. & Provost, F. Deep learning on big, sparse, behavioral data. *Big Data* **7**, 286–307 (2019).
52. Kirchhof, M., Mucsányi, B., Oh, S. J. & Kasneci, D. E. Url: A representation learning benchmark for transferable uncertainty estimates. *Adv. Neural Inf. Process. Syst.* **36**, 13956–13980 (2023).
53. Upadhyay, U., Kim, J. M., Schmidt, C., Schölkopf, B. & Akata, Z. Posterior annealing: Fast calibrated uncertainty for regression. Preprint at https://arxiv.org/abs/2302.11012 (2023).
54. Ashburner, M. et al. Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
55. Aleksander, S. A. et al. The gene ontology knowledgebase in 2023. *Genetics* **224**, iyad031 (2023).
56. Kim-Hellmuth, S. et al. Cell type–specific genetic regulation of gene expression across human tissues. *Science* **369**, eaaz8528 (2020).
57. Hagai, T. et al. Gene expression variability across cells and species shapes innate immunity. *Nature* **563**, 197–202 (2018).
58. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res.* **25**, 1491–1498 (2015).
59. Ng, P. C. & Kirkness, E. F. *Genetic variation: Methods and Protocols* 215–226 (2010).
60. Consortium, G. The gtex consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
61. Sasse, A. et al. Benchmarking of deep neural networks for predicting personal gene expression from dna sequence highlights shortcomings. *Nat. Genet.* **55**, 2060–2064 (2023).
62. Huang, C. et al. Personal transcriptome variation is poorly explained by current genomic deep learning models. *Nat. Genet.* **55**, 2056–2059 (2023).
63. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).
64. Das, B. K. et al. The cytoplasmic dynein associated protein nde1 regulates osteoclastogenesis by modulating m-csf and rankl signaling pathways. *Cells* **11**, 13 (2021).
65. Li, Y. et al. Single-cell transcriptome analysis reveals dynamic cell populations and differential gene expression patterns in control and aneurysmal human aortic tissue. *Circulation* **142**, 1374–1388 (2020).
66. Chaffin, M. et al. Single-nucleus profiling of human dilated and hypertrophic cardiomyopathy. *Nature* **608**, 174–180 (2022).
67. Szałata, A. et al. A benchmark for prediction of transcriptomic responses to chemical perturbations across cell types. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (2024).
68. Dann, E. et al. Precise identification of cell states altered in disease using healthy single-cell references. *Nat. Genet.* **55**, 1998–2008 (2023).
69. Chen, X., Zhang, Z., Wang, X., Chen, Y. & Wang, C. Nt5c2 gene polymorphisms and the risk of coronary heart disease. *Public Health Genomics* **23**, 90–99 (2020).
70. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
71. Yeung, W., Zhou, Z., Li, S. & Kannan, N. Alignment-free estimation of sequence conservation for identifying functional sites using protein sequence embeddings. *Brief. Bioinform.* **24**, bbac599 (2023).
72. Saraswat, D. et al. Explainable ai for healthcare 5.0: opportunities and challenges. *IEEe Access* **10**, 84486–84517 (2022).
73. Yang, J., Lindenbaum, O. & Kluger, Y. Locally sparse neural networks for tabular biomedical data. In *Proceedings International Conference on Machine Learning*, 25123–25153 (PMLR, 2022).
74. Chaddad, A., Peng, J., Xu, J. & Bouridane, A. Survey of explainable ai techniques in healthcare. *Sensors* **23**, 634 (2023).
75. Wysocka, M., Wysocki, O., Zufferey, M., Landers, D. & Freitas, A. A systematic review of biologically-informed deep learning models

for cancer: fundamental trends for encoding and interpreting oncology data. *BMC Bioinform.* **24**, 198 (2023).
76. He, Y. et al. Robust multi-task learning with excess risks. In *Proceedings of the 41st International Conference on Machine Learning*, 18094–18114 (2024).
77. Zhang, Y. & Yang, Q. An overview of multi-task learning. *Natl. Sci. Rev.* **5**, 30–43 (2018).
78. Cao, Z.-J. & Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.* **40**, 1458–1466 (2022).
79. Theodoris, C. V. et al. Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).
80. Hoang, D.-T. et al. A deep-learning framework to predict cancer treatment response from histopathology images through imputed transcriptomics. *Nat. Cancer* **5**, 1305–1317 (2024).
81. Wang, C. et al. Benchmarking the translational potential of spatial gene expression prediction from histology. *Nat. Commun.* **16**, 1544 (2025).
82. Chung, Y., Ha, J. H., Im, K. C. & Lee, J. S. Accurate spatial gene expression prediction by integrating multi-resolution features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11591–11600 (2024).
83. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
84. Wolf, F. A., Angerer, P. & Theis, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 1–5 (2018).
85. HelloWorldLTY. Helloworldlty/unicorn: Release zendo file. https://doi.org/10.5281/zenodo.16416845 (2025).
86. JGraph. diagrams.net, draw.io https://www.diagrams.net/ (2021).

Acknowledgements

We appreciate the comments, feedback, and suggestions from Dr. Gokcen Eraslan for model training. This project is supported in part by NIH grants U01 HG013840, U24 HG012108, and P50 CA196530.

Author contributions

T.L. and Y.L. designed this study. T.L. and T.H. designed the model. T.L., T.H., and L.W. ran all the experiments. T.L., T.H., L.W., Y.L., R.Y., and H.Z. wrote the manuscript. R.Y. and H.Z. provided the computation resources. H.Z. supervised this work.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-64506-8.

Correspondence and requests for materials should be addressed to Hongyu Zhao.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025