UNICORN: Towards Universal Cellular Expression Prediction with an Explainable Multi-Task Learning Framework

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

[1]Interdepartmental Program in Computational Biology & Bioinformatics, Yale University, New Haven, 06511, CT, USA.
[2]Department of Biostatistics, Yale University, New Haven, 06511, CT, USA.
[3]Department of Computer Science, Yale University, New Haven, 06511, CT, USA.

*Corresponding author(s). E-mail(s): hongyu.zhao@yale.edu;
Contributing authors: tianyu.liu@yale.edu; tinglin.huang@yale.edu; yingxin.lin@yale.edu; rex.ying@yale.edu;

Abstract

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 new 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.

Keywords: DNA Sequence Model, Large Language Model, Single-Cell Sequencing, Multi-Modal Machine Learning, Multi-Task Machine Learning

1 Introduction

Modelling DNA and 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 measurement 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 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]) 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 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 a novel 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**edictio**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.

2 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 in a separate file. We then construct the training and testing datasets based on the target omic type and train multiple non-linear predictors to model the relationship between sequence and expression information. Meanwhile, we also have an uncertainty estimator to quantify the uncertainty of each prediction. 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, and disease-specific levels. The landscape of UNICORN is summarized in Figures 1 (a)-(c).

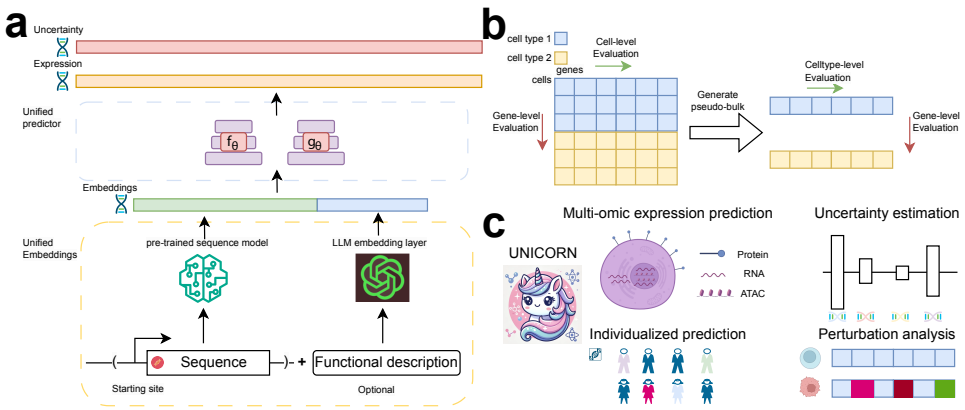


Fig. 1 Workflow and functions of UNICORN. (a) The architecture and components of UNICORN. Our method accepts embeddings from multi-modal sources, predicts expression levels at the single-cell or cell-type resolutions, and estimates the uncertainty of model outputs. (b) Explanations of our evaluation 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 logo is generated with DALLE [35].

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 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, 36]. 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), UNICORN with gene embeddings from HyenaDNA (UNICORN_hyenadna), Borzoi (UNICORN_borzoi), Nucleotide Transformers [37] (UNICORN_nt), DNABERT (UNICORN_dnabert), UNICORN with gene embeddings from Gene2vec [38] (UNICORN_gene2vec), UNICORN with gene embeddings from LLMs [39](UNICORN_genept), and UNICORN with 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.

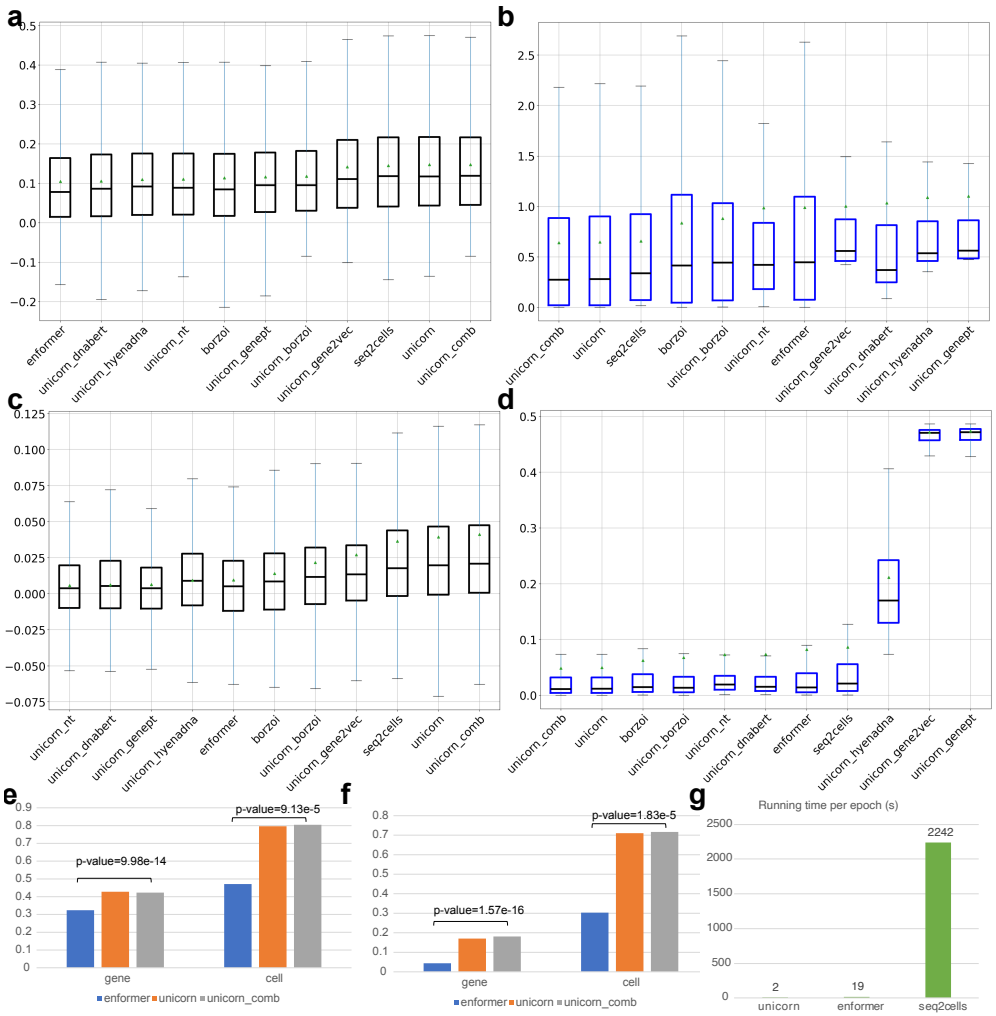


Fig. 2 Comprehensive evaluations for gene expression predictions from sequences. The triangle shape represents the mean value, and the black dashed line represents the median value. We report the results based on box plots. (a) Results of gene-level correlations across different methods based on the thymus dataset. (b) Results of MSE across different methods based on the thymus dataset. (c) Results of gene-level correlations across different methods based on the PBMC dataset. (d) Results of MSE across different methods based on the PBMC dataset. (e) Results of gene-level correlations with pseudo-bulk evaluation based on the thymus dataset. (f) Results of gene-level correlations with pseudo-bulk evaluation based on the PBMC dataset. (g) Training time per epoch of different models. The unit is second (s).

We present the prediction results at the single-cell resolution in Figures 2 (a)-(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 2 (a) 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 smaller averaged MSE according to Figure 2 (b), 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 Figures 2 (c) and (d). For this dataset, UNICORN and UNICORN_combine had high correlation scores while UNICORN had the lowest MSE. We found that UNICORN with gene2vec, genept, and combined information as sources of gene embeddings had 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 Extended Data Figures 1 (a)-(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_corr, shown in Extended Data Figure 2. 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 [40]. In conclusion, our experimental results demonstrate that sequence information can serve as a better starting point for analyzing gene expression at the cellular level.

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

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 Extended Data Figure 3, 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 Extended Data Figures 4 (a) and (b), for these two datasets. We investigated the selected sequence length's effect on prediction performances, summarized in Appendix A. Our model can also predict gene expression levels for atlas-level datasets, which is further discussed in Appendix B.

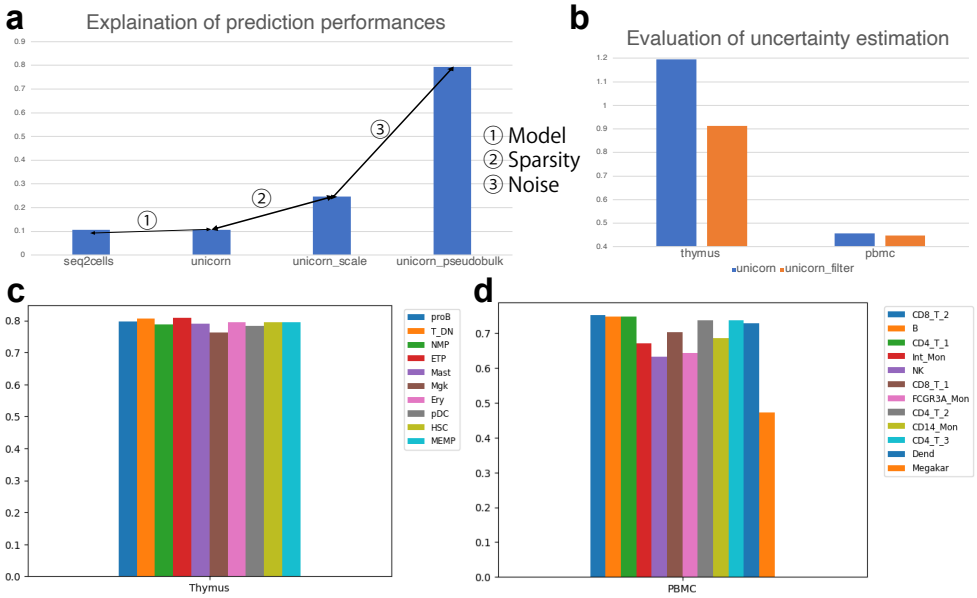


Fig. 3 Explainability of UNICORN helps understand the principles of expression prediction. (a) Explainability of UNICORN from the differences of pre-processing steps and post-processing steps. (b) Explainability of UNICORN from uncertainty estimation. (c) Average correlation coefficients across different cell types for the thymus dataset. (d) Average correlation coefficients across different cell types for the PBMC dataset.

Uncovering the complexity of model training. One significant contribution of UNICORN is to generalize an explainable framework for expression prediction. Here, we discuss the first component of explainability, which helps us understand the behavior of gene transcription prediction. In Figure 3 (a) 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 hyper-parameters and test a better loss function design. Such 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 [41] to impute the scRNA-seq dataset as a denoising step. However, such design is biased to the imputation method because we found that using different denoising methods led to different results, shown in Extended Data Figure 5 including the comparison between scVI and DeepImpute [42]. A more general approach used in single-cell data analysis to reduce the sparsity is to perform scaling, so we scaled the single-cell dataset using

z-norm for each gene. Such 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 extremely 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 Figure 3 (a). 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 Appendix C.

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 [43]. Our idea is to estimate the uncertainty levels of each sample to identify the reliability of model prediction. We computed the uncertainty of each gene and each cell based on the loss prediction method (LossPred) [44]. As shown in Figure 3 (b), 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 Extended Data Figure 6. 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) [45, 46] with the enriched pathways shown in Extended Data Figures 7 (a) and (b). 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, 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. These pathways work to maintain basic life activities, and the specific parts of each gene set may suggest tissue-specific activities. In conclusion, our uncertainty estimation component can help us identify difficult components of expression prediction.

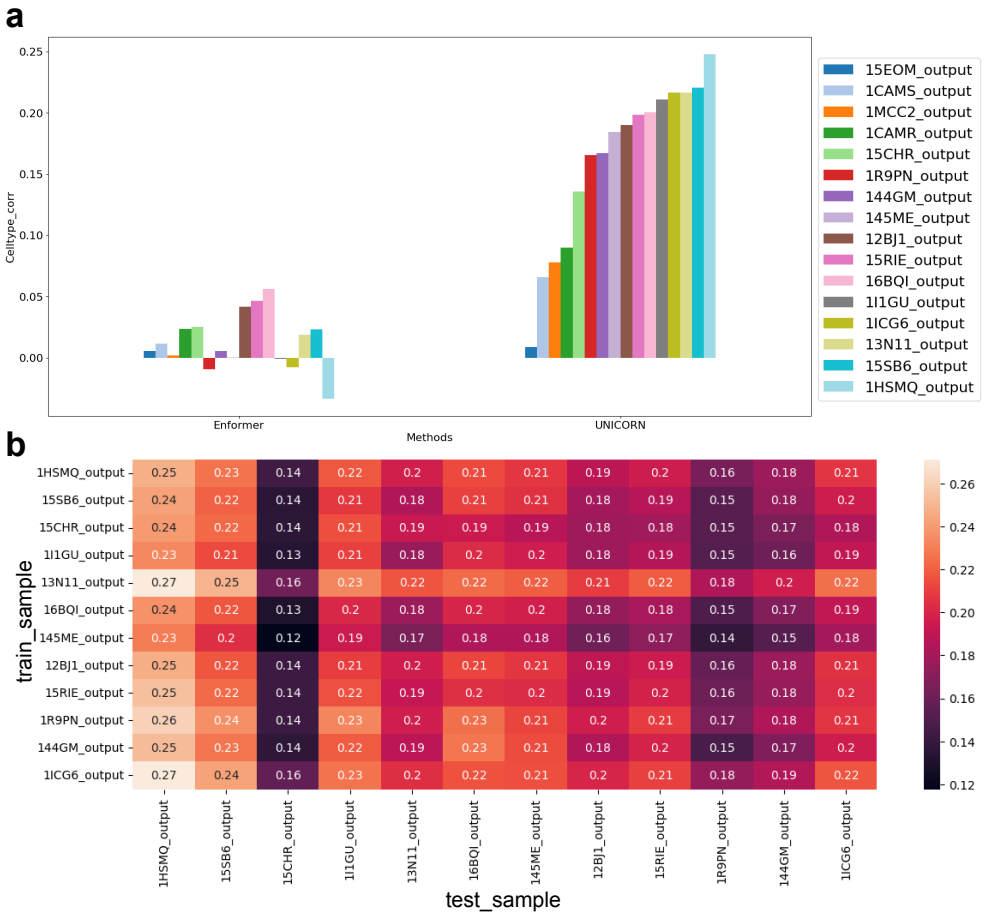


Fig. 4 Results of gene expression predictions for individualized scRNA-seq datasets. The shared cell type is endothelial cell. (a) The results of cell-type correlations computed based on the given cell type across individuals for outputs from Enformer and UNICORN. Samples are ranked based on the cell-type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

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 [47–49]. 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 Figure 3 (c) for thymus and in Figure 3 (d) 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) [50] 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 [51]. 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 Extended Data Figure 9, 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 Figure 4 (a). Our findings also matched previous conclusions on the drawbacks of using Enformer for individualized gene expression prediction [52, 53]. 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 visualize the comparison between fine-tuned Enformer and UNICORN in Extended Data Figure 8 (a). 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 Extended Data Figures 10 (a) and (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 given datasets, shown in Extended Data Figure 8 (b). 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 Extended Data Figures 8 (c) and (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 Figure 4 (a). 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 Figure 4 (b) and Extended Data Figure 10 (c). 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 [54]. Our discussion can be found in Appendix F. 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.

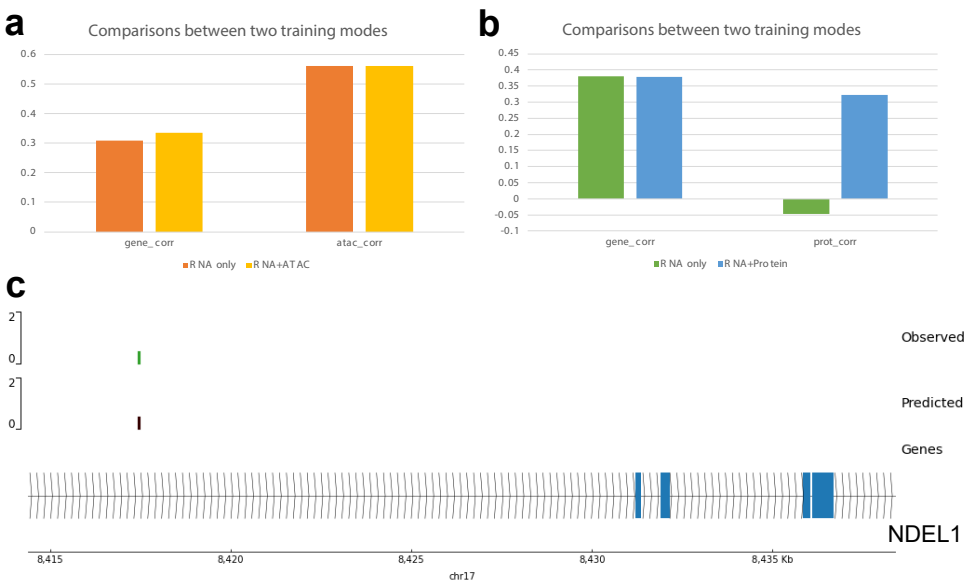


Fig. 5 UNICORN demonstrates its ability in predicting multi-omic information at the cell-type level. Here RNA represents the gene expression profiles. The evaluation is performed based on the pseudo-bulk data. We reported the averaged correlation scores in (a) and (b). (a) The comparisons between RNA 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.

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 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 Figure 5 (a) for the 10x Multi-omic dataset and Figure 5 (b) 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 Extended Data Figure 11 (a). 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 Extended Data Figure 11 (b). Furthermore, this gene overlaps with NDEL1, as shown in Figure 5 (c). NDEL1 expressed in monocytes has proved to be able to regulate osteoclasts and bone homeostasis [55], and thus UNICORN can accurately predict the important peak associated with biological functions.

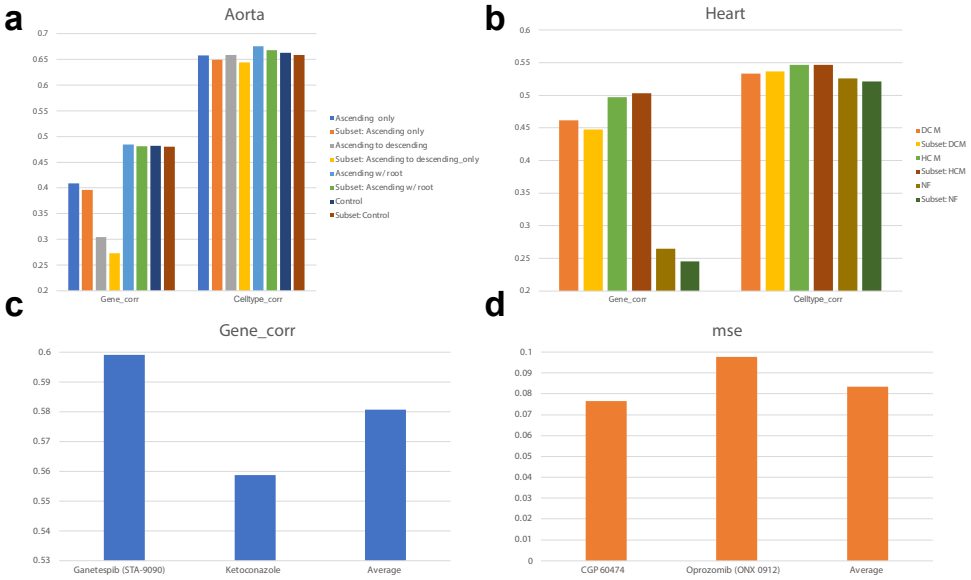


Fig. 6 UNICORN uncovers the complexity of diseases and perturbations. (a) The correlation coefficients computed based on genes and cells across different conditions from the Aorta dataset. (b) The correlation coefficients computed based on genes and cells across different conditions from the Heart dataset. (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.

Predicting disease effects and perturbation effects uncovers the biological variation. By performing prediction for gene expression levels under different conditions (including diseased/non-diseased and perturbed/non-perturbed), UNICORN can also help us understand the biological 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 [56] and Heart [57] are used to investigate the effects of diseases, and the dataset Openproblem [58] 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 hyper-parameters 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 6 (a) 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 6 (b) shows the prediction results based on the Heart dataset, and the labeled results like "Subset: NF" represents 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 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 [59] 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. Therefore, UNICORN can help us analyze the differences of various conditions and suggests the importance of collecting cells from different states to build atlas-level databases.

We then investigated the performances of UNICORN in predicting gene expression levels under different perturbations. Figure 6 (c) 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 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 high correlation score (gene_corr=0.70) and the DEG MX1 of Ketoconazole had low correlation score (gene_corr=-0.77). Figure 6 (d) 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.

3 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 regulatory elements and gene expressions at the cell-type or cell-state levels. In this work, we present a novel framework, called 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 novel 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 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. 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 a novel 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.

4 Methods

Problem statement. Considering we have a single-cell dataset $\mathcal{D}$ with modalities $1...m$, we then can denote the single-cell dataset as $\mathcal{D} = \{X_1^{n\times p_1}, X_2^{n\times p_2}, ..., X_m^{n\times p_m}\}$. Each expression matrix X has the same number of sequenced cells n but a different number of biological features $p_1...p_m$. 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 $\mathcal{P} = \{e_1^{d_1\times p_1}, e_2^{d_2\times p_2}, ..., e_2^{d_m\times p_m}\}$. 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 θ_1 and θ_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 ith modality as an example.

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 target datasets determine 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 [60] 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 [61]. 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 [36] as the activation function and Softplus [36] 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. Our estimator and predictor have the same architectures but do not share weights because they work for different predictions. 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. The method we used to estimate the uncertainty for prediction output is LossPred, which means we directly predict the MSE of the given training example. Such a method, supported by [43], outperformed other methods in their evaluation results.

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:

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

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 ith cell and j^{th} feature in the predicted matrix. The weights of these loss components are the same. To elaborate on the necessity of including all three loss components to formalize our final loss function, we perform ablation tests shown in Extended Data Figures 12 (a)-(c) under different metrics.

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. Since data from different modalities may follow different distributions, we only utilized the PCC loss as the loss function. We demonstrated the contributions of multi-modal prediction by comparing it with prediction based on a single modality.

We also discuss the hyper-parameter 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 Appendix D.

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 [62] 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 [63]. These genes are classified into protein-encoding genes, non-coding RNA, etc. Details of gene functional list can be found in Extended Data Figure 15 (a).

To infer the functions of gene sets we are interested in, we utilize Gene Ontology Enrichment Analysis (GOEA) [45, 46] 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. Our metrics are computed using the scikit-learn package [64] 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 performances. The value of MSE is in $(0, \infty)$, and a lower value represents better model performances.

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 seq2cells [11], Enformer [21], Borzoi [24], Decima [22], and variations of UNICORN as baselines. seq2cells is a model based on the sequence embeddings from pre-trained Enformer to predict

gene expression levels in 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. 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) [39], embeddings from the DNA foundation model (HyenaDNA [2], DNABERT [29]), and embeddings from gene-gene interaction (gene2vec) [38], as inputs.

For the task of model explainability, we consider scVI [41] and DeepImpute [42] 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 [65]. We compared the performances of UNICORN between the raw mode and normalized mode in Appendix E. 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.

Data availability. We include the links to access our experimental datasets in Supplementary file 1. For the datasets from GTEx, users need to purchase this database for access. For other datasets, they are publicly available.

Device and codes information. 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 analysis is 24 hours.

The codes of UNICORN can be found in https://github.com/HelloWorldLTY/UNICORN. The license is MIT license.

5 Ethics and Inclusion.

Although UNICORN is not biased to gender, races, and other factors, the users are solely responsible for the content they generate with models in UNICORN, and there are no mechanisms in place for addressing harmful, unfaithful, biased, and toxic content disclosure. Any modifications of the models should be released under different version numbers to keep track of the original models related to this manuscript.

The target of current UNICORN only serves for academic research. The users cannot use it for other purposes. Finally, we are not responsible for any effects of the use of the model.

6 Acknowledgments

We appreciate the comments, feedback, and suggestions from Dr. Gokcen Eraslan for model training.

7 Author contributions

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

References

[1] Yang, Z.: A space-time process model for the evolution of dna sequences. Genetics **139**(2), 993–1005 (1995)

[2] Nguyen, E., Poli, M., Faizi, M., Thomas, A., Wornow, M., Birch-Sykes, C., Massaroli, S., Patel, A., Rabideau, C., Bengio, Y., 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 Sciences Education **15**(1), 4 (2016)

[4] Zhou, J., Troyanskaya, O.G.: Predicting effects of noncoding variants with deep learning–based sequence model. Nature methods **12**(10), 931–934 (2015)

[5] Zhou, J., Theesfeld, C.L., Yao, K., Chen, K.M., Wong, A.K., Troyanskaya, O.G.: Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nature genetics **50**(8), 1171–1179 (2018)

[6] Ramprasad, P., Pai, N., Pan, W.: Enhancing personalized gene expression prediction from dna sequences using genomic foundation models. Human Genetics and Genomics Advances (2024)

[7] Drusinsky, S., Whalen, S., Pollard, K.S.: Deep-learning prediction of gene expression from personal genomes. bioRxiv, 2024–07 (2024)

[8] Beer, M.A., Tavazoie, S.: Predicting gene expression from sequence. Cell **117**(2), 185–198 (2004)

[9] Clark, W.T., Radivojac, P.: Analysis of protein function and its prediction from amino acid sequence. Proteins: Structure, Function, and Bioinformatics **79**(7), 2086–2096 (2011)

[10] Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A.J., Bambrick, J., et al.: Accurate structure

prediction of biomolecular interactions with alphafold 3. Nature, 1–3 (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. bioRxiv, 2023–07 (2023)

[12] Han, X., Zhou, Z., Fei, L., Sun, H., Wang, R., Chen, Y., Chen, H., Wang, J., Tang, H., Ge, W., *et al.*: Construction of a human cell landscape at single-cell level. Nature **581**(7808), 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 research **42**(14), 8845–8860 (2014)

[14] Lance, C., Luecken, M.D., Burkhardt, D.B., Cannoodt, R., Rautenstrauch, P., Laddach, A., Ubingazhibov, A., Cao, Z.-J., Deng, K., Khan, S., Liu, Q., Russkikh, N., Ryazantsev, G., Ohler, U., participants, N..M., Pisco, A.O., Bloom, J., Krishnaswamy, S., Theis, F.J.: Multimodal single cell data integration challenge: Results and lessons learned. In: Kiela, D., Ciccone, M., Caputo, B. (eds.) Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track. Proceedings of Machine Learning Research, vol. 176, pp. 162–176. PMLR, ??? (2022). https://proceedings.mlr.press/v176/lance22a.html

[15] Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., Greenleaf, W.J.: Single-cell chromatin accessibility reveals principles of regulatory variation. Nature **523**(7561), 486–490 (2015)

[16] Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P.K., Swerdlow, H., Satija, R., Smibert, P.: Simultaneous epitope and transcriptome measurement in single cells. Nature methods **14**(9), 865–868 (2017)

[17] Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., *et al.*: Accounting for technical noise in single-cell rna-seq experiments. Nature methods **10**(11), 1093–1095 (2013)

[18] You, Y., Dong, X., Wee, Y.K., Maxwell, M.J., Alhamdoosh, M., Smyth, G.K., Hickey, P.F., Ritchie, M.E., Law, C.W.: Modeling group heteroscedasticity in single-cell rna-seq pseudo-bulk data. Genome biology **24**(1), 107 (2023)

[19] Nathan, A., Asgari, S., Ishigaki, K., Valencia, C., Amariuta, T., Luo, Y., Beynor, J.I., Baglaenko, Y., Suliman, S., Price, A.L., *et al.*: Single-cell eqtl models reveal dynamic t cell state dependence of disease loci. Nature **606**(7912), 120–128 (2022)

[20] Liu, T., Long, W., Cao, Z., Wang, Y., He, C.H., Zhang, L., Strittmatter, S.M., Zhao, H.: Cosgenegate selects multi-functional and credible biomarkers for single-cell analysis. Briefings in Bioinformatics **26**(1), 626 (2025)

[21] Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J.R., Grabska-Barwinska, A., Taylor, K.R., Assael, Y., Jumper, J., Kohli, P., Kelley, D.R.: Effective gene expression prediction from sequence by integrating long-range interactions. Nature methods **18**(10), 1196–1203 (2021)

[22] Lal, A., Karollus, A., Gunsalus, L., Garfield, D., Nair, S., Tseng, A.M., Gordon, M.G., Collier, J.L., Diamant, N., Biancalani, T., et al.: Decoding sequence determinants of gene expression in diverse cellular and disease states. bioRxiv, 2024–10 (2024)

[23] Kelley, D.R., Reshef, Y.A., Bileschi, M., Belanger, D., McLean, C.Y., Snoek, J.: Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome research **28**(5), 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. bioRxiv, 2023–08 (2023)

[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 medicine **9**, 1–12 (2017)

[26] Li, J., Wang, J., Zhang, P., Wang, R., Mei, Y., Sun, Z., Fei, L., Jiang, M., Ma, L., E, W., *et al.*: Deep learning of cross-species single-cell landscapes identifies conserved regulatory programs underlying cell types. Nature Genetics **54**(11), 1711–1720 (2022)

[27] Hingerl, J.C., Martens, L.D., Karollus, A., Manz, T., Buenrostro, J.D., Theis, F.J., Gagneur, J.: scooby: Modeling multi-modal genomic profiles from dna sequence at single-cell resolution. bioRxiv, 2024–09 (2024)

[28] Jaiswal, A., Babu, A.R., Zadeh, M.Z., Banerjee, D., Makedon, F.: A survey on contrastive self-supervised learning. Technologies **9**(1), 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**(15), 2112–2120 (2021)

[30] Nguyen, E., Poli, M., Durrant, M.G., Kang, B., Katrekar, D., Li, D.B., Bartie, L.J., Thomas, A.W., King, S.H., Brixi, G., *et al.*: Sequence modeling and design from molecular to genome scale with evo. Science **386**(6723), 9336 (2024)

[31] Marin, F.I., Teufel, F., Horlacher, M., Madsen, D., Pultz, D., Winther, O., Boomsma, W.: 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 systems **12**(6), 654–669 (2021)

[33] Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al.: A survey of large language models. arXiv preprint arXiv:2303.18223 (2023)

[34] Nachtomy, O., Shavit, A., Yakhini, Z.: Gene expression and the concept of the phenotype. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences **38**(1), 238–254 (2007)

[35] OpenAI: DALLE. OpenAI (2024)

[36] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)

[37] Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Lopez Carranza, N., Grzywaczewski, A.H., Oteri, F., Dallago, C., Trop, E., Almeida, B.P., Sirelkhatim, H., et al.: Nucleotide transformer: building and evaluating robust foundation models for human genomics. Nature Methods, 1–11 (2024)

[38] Du, J., Jia, P., Dai, Y., Tao, C., Zhao, Z., Zhi, D.: Gene2vec: distributed representation of genes based on co-expression. BMC genomics **20**, 7–15 (2019)

[39] Chen, Y., Zou, J.: Genept: A simple but effective foundation model for genes and cells built from chatgpt. bioRxiv (2023)

[40] Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y.: Decoupling representation and classifier for long-tailed recognition. In: International Conference on Learning Representations (2020). https://openreview.net/forum?id=r1gRTCVFvB

[41] Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., Yosef, N.: Deep generative modeling for single-cell transcriptomics. Nature methods **15**(12), 1053–1058 (2018)

[42] 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 biology **20**, 1–14 (2019)

[43] Kirchhof, M., Mucsányi, B., Oh, S.J., Kasneci, D.E.: Url: A representation learning benchmark for transferable uncertainty estimates. Advances in Neural Information Processing Systems **36**, 13956–13980 (2023)

[44] Upadhyay, U., Kim, J.M., Schmidt, C., Schölkopf, B., Akata, Z.: Posterior

annealing: Fast calibrated uncertainty for regression. arXiv e-prints, 2302 (2023)

[45] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., *et al.*: Gene ontology: tool for the unification of biology. Nature genetics **25**(1), 25–29 (2000)

[46] Aleksander, S.A., Balhoff, J., Carbon, S., Cherry, J.M., Drabkin, H.J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N.L., *et al.*: The gene ontology knowledgebase in 2023. Genetics **224**(1), 031 (2023)

[47] Kim-Hellmuth, S., Aguet, F., Oliva, M., Muñoz-Aguirre, M., Kasela, S., Wucher, V., Castel, S.E., Hamel, A.R., Viñuela, A., Roberts, A.L., *et al.*: Cell type–specific genetic regulation of gene expression across human tissues. Science **369**(6509), 8528 (2020)

[48] Hagai, T., Chen, X., Miragaia, R.J., Rostom, R., Gomes, T., Kunowska, N., Henriksson, J., Park, J.-E., Proserpio, V., Donati, G., *et al.*: Gene expression variability across cells and species shapes innate immunity. Nature **563**(7730), 197–202 (2018)

[49] Trapnell, C.: Defining cell types and states with single-cell genomics. Genome research **25**(10), 1491–1498 (2015)

[50] Ng, P.C., Kirkness, E.F.: Whole genome sequencing. Genetic variation: Methods and protocols, 215–226 (2010)

[51] Consortium, G.: The gtex consortium atlas of genetic regulatory effects across human tissues. Science **369**(6509), 1318–1330 (2020)

[52] Sasse, A., Ng, B., Spiro, A.E., Tasaki, S., Bennett, D.A., Gaiteri, C., De Jager, P.L., Chikina, M., Mostafavi, S.: Benchmarking of deep neural networks for predicting personal gene expression from dna sequence highlights shortcomings. Nature Genetics **55**(12), 2060–2064 (2023)

[53] Huang, C., Shuai, R.W., Baokar, P., Chung, R., Rastogi, R., Kathail, P., Ioannidis, N.M.: Personal transcriptome variation is poorly explained by current genomic deep learning models. Nature Genetics **55**(12), 2056–2059 (2023)

[54] Schaid, D.J., Chen, W., Larson, N.B.: From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics **19**(8), 491–504 (2018)

[55] Das, B.K., Gogoi, J., Kannan, A., Gao, L., Xing, W., Mohan, S., Zhao, H.: The cytoplasmic dynein associated protein nde1 regulates osteoclastogenesis by modulating m-csf and rankl signaling pathways. Cells **11**(1), 13 (2021)

[56] Li, Y., Ren, P., Dawson, A., Vasquez, H.G., Ageedi, W., Zhang, C., Luo, W., Chen,

R., Li, Y., Kim, S., *et al.*: Single-cell transcriptome analysis reveals dynamic cell populations and differential gene expression patterns in control and aneurysmal human aortic tissue. Circulation **142**(14), 1374–1388 (2020)

[57] Chaffin, M., Papangeli, I., Simonson, B., Akkad, A.-D., Hill, M.C., Arduini, A., Fleming, S.J., Melanson, M., Hayat, S., Kost-Alimova, M., *et al.*: Single-nucleus profiling of human dilated and hypertrophic cardiomyopathy. Nature **608**(7921), 174–180 (2022)

[58] Szałata, A., Benz, A., Cannoodt, R., Cortes, M., Fong, J., Kuppasani, S., Lieberman, R., Liu, T., Mas-Rosario, J.A., Meinl, R., Nourisa, J., Tumiel, J., Tunjic, T.M., Wang, M., Weber, N., Zhao, H., Anchang, B., Theis, F.J., Luecken, M.D., Burkhardt, D.B.: 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). https://openreview.net/forum?id=WTI4RJYSVm

[59] Dann, E., Cujba, A.-M., Oliver, A.J., Meyer, K.B., Teichmann, S.A., Marioni, J.C.: Precise identification of cell states altered in disease using healthy single-cell references. Nature Genetics **55**(11), 1998–2008 (2023)

[60] Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., *et al.*: Evolutionary-scale prediction of atomic-level protein structure with a language model. Science **379**(6637), 1123–1130 (2023)

[61] Yeung, W., Zhou, Z., Li, S., Kannan, N.: Alignment-free estimation of sequence conservation for identifying functional sites using protein sequence embeddings. Briefings in Bioinformatics **24**(1), 599 (2023)

[62] Cao, Z.-J., Gao, G.: Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. Nature Biotechnology **40**(10), 1458–1466 (2022)

[63] Theodoris, C.V., Xiao, L., Chopra, A., Chaffin, M.D., Al Sayed, Z.R., Hill, M.C., Mantineo, H., Brydon, E.M., Zeng, Z., Liu, X.S., *et al.*: Transfer learning enables predictions in network biology. Nature **618**(7965), 616–624 (2023)

[64] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., *et al.*: Scikit-learn: Machine learning in python. the Journal of machine Learning research **12**, 2825–2830 (2011)

[65] Wolf, F.A., Angerer, P., Theis, F.J.: Scanpy: large-scale single-cell gene expression data analysis. Genome biology **19**, 1–5 (2018)

[66] Luecken, M.D., Theis, F.J.: Current best practices in single-cell rna-seq analysis: a tutorial. Molecular systems biology **15**(6), 8746 (2019)

[67] Ahlmann-Eltze, C., Huber, W.: Comparison of transformations for single-cell rna-seq data. Nature Methods **20**(5), 665–672 (2023)

[68] Zhang, M.J., Ntranos, V., Tse, D.: Determining sequencing depth in a single-cell rna-seq experiment. Nature communications **11**(1), 774 (2020)

[69] Hafemeister, C., Satija, R.: Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. Genome biology **20**(1), 296 (2019)

[70] Wang, J., Huang, D., Zhou, Y., Yao, H., Liu, H., Zhai, S., Wu, C., Zheng, Z., Zhao, K., Wang, Z., *et al.*: Causaldb: a database for disease/trait causal variants identified using summary statistics of genome-wide association studies. Nucleic acids research **48**(D1), 807–816 (2020)

[71] Fagny, M., Paulson, J.N., Kuijjer, M.L., Sonawane, A.R., Chen, C.-Y., Lopes-Ramos, C.M., Glass, K., Quackenbush, J., Platig, J.: Exploring regulation in tissues with eqtl networks. Proceedings of the National Academy of Sciences **114**(37), 7841–7850 (2017)

[72] Ramos, E.M., Din-Lovinescu, C., Berg, J.S., Brooks, L.D., Duncanson, A., Dunn, M., Good, P., Hubbard, T.J., Jarvik, G.P., O'Donnell, C., *et al.*: Characterizing genetic variants for clinical action. In: American Journal of Medical Genetics Part C: Seminars in Medical Genetics, vol. 166, pp. 93–104 (2014). Wiley Online Library

[73] Su, J., Han, C., Zhou, Y., Shan, J., Zhou, X., Yuan, F.: Saprot: Protein language modeling with structure-aware vocabulary. In: The Twelfth International Conference on Learning Representations (2024). https://openreview.net/forum?id=6MRm3G4NiU

A How to describe the effect of input sequence length to gene expression prediction performances?

In this section, we intend to investigate the contributions of increasing window size of sequence length for gene expression prediction. Previous methods claimed that sequence-to-function models should capture long context to enhance more performances, so we hypothesized that such a conclusion should also fit this problem. To validate our assumption without being biased by the design of pre-trained models, we paired the sequence information centered by TSS but with different context lengths with single-cell gene expression profiles, and trained a linear regression model to evaluate our results. According to Extended Data Figure 13, the MSE increased as we increased the length of the context window, while the other two correlation-based scores also increased. Therefore, expanding the context window helps in learning the correlation between prediction outputs and observed gene expression levels, which might not help in learning the specific magnitude of gene expression levels.

B Can we predict gene expression levels in atlas-level single-cell datasets?

In this section, we demonstrate that UNICORN is capable of predicting gene expression levels in the atlas dataset from human thymus (thymus_atlas). Our dataset contains 255,901 cells and 18,204 genes. We used the same training setting and generated the gene embeddings using PCA from the observed gene expression and predicted gene expression for training genes shown in Extended Data Figure 14 (a) and testing genes shown in Extended Data Figure 14 (b). This figure shows that our predicted gene embeddings overlap with observed gene embeddings well for most of genes. Meanwhile, for the genes whose embeddings are not aligned well under the observed case and the predicted case, the observed gene embeddings show clustering patterns which are less informative under both training and testing cases. Therefore, the unaligned patterns of gene embeddings might come from the quality of data or genes rather than the capacity of UNICORN. Moreover, we also visualize the distribution of cells based on the predicted gene expression levels in Extended Data Figure 14 (c), which preserves the separation of different cell types in a good manner. Therefore, our method also supports atlas-level prediction without facing memory issues.

C Can the functional annotations of genes fully explain our prediction performances?

In this section, we collect functional labels (e.g., protein-encoding and lincRNA) of genes and use different gene embeddings to quantify them. Our evaluation framework is inspired by the difference of their functionalities and the clustering performance based on gene embeddings can reflect the ability to represent gene functions. Understanding the relationship between gene functions and gene embeddings may help us explore

the contributions of gene embeddings on expression prediction. In total we consider gene embeddings from GenePT, Gene2vec, Enformer, Enformer+GenePT (Combine), HyenaDNA and DNABERT and benchmark their performances on gene clustering. The UMAP plots of different gene embeddings are shown in Extended Data Figure 15 (a), with all of them being similar because the dominant genes are protein-encoding genes for scRNA-seq datasets. Furthermore, we compute three different metrics for evaluating clustering performances, including Normalized Mutual Information (NMI), Adjusted Rand Index (ARI) and Averaged silhouette Width, and average them to have the final score. We rank these method based on this score, shown in Extended Data Figure 15 (b). This figure shows that the best model for gene clustering is GenePT, followed by Gene2vec and Enformer. This rank is not highly correlated to the method's rank for gene expression prediction. Therefore, our discovery might explain why direct inference based on gene embeddings is difficult for expression prediction, and further supports the utilization of our adapter design.

D Do we need to tune hyper-parameters for prediction?

In this section, we discuss the hyper-parameter tuning process for UNICORN based on the validation dataset. Our best epoch is determined by early-stopping with the help of a validation dataset. Therefore, we only considered the effects of 1. Learning Rate (LR) and 2. Dropout Rate (DR). Our results are summarized in Extended Data Figure 16 (a) for LR and Extended Data Figure 16 (b) for DR. We found that LR has larger effects for model performances and we finally set LR as 1e-4. Furthermore, DR does not affect model performances obviously and we finally set DR as 0.4. Such setting is kept the same across all of the experiments in this manuscript.

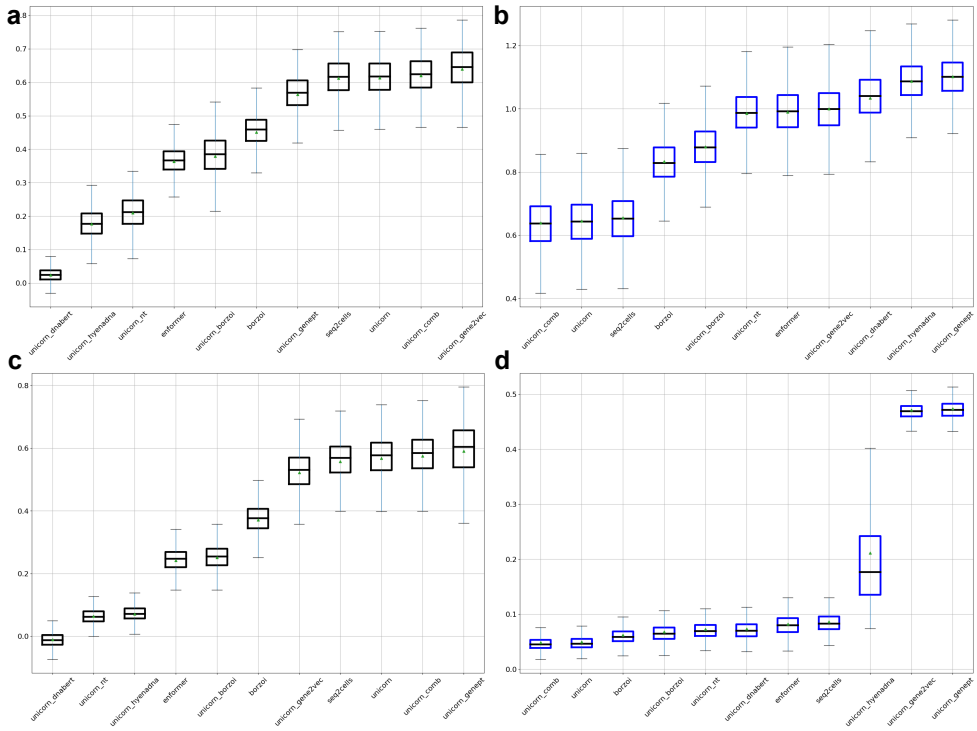
E Do we need normalized single-cell data for prediction?

The normalization approach for pre-processing single-cell data has been well-discussed [66, 67], and the raw single-cell data are biased to the sequence depth [68]. To validate the necessity of utilization log-norm data for prediction, we utilize the PBMC dataset as an example and perform prediction for raw count data, log-normalized data, and sctransform [69] data as targets. Our results are shown in Extended Data Figures 17 (a)-(d). According to these figures, prediction outcomes based on log-normalized data have lower variance and MSE, and its correlation coefficients based on cells are also higher than the raw data mode, while comparable with the sctransform mode. Prediction results based on sctransform data also have higher MSE than the log-normalized mode result, which suggest that the log-normalized approach is more suitable in this case. Therefore, the utilization of log-normalized data can help us generate more stable and reliable outputs.

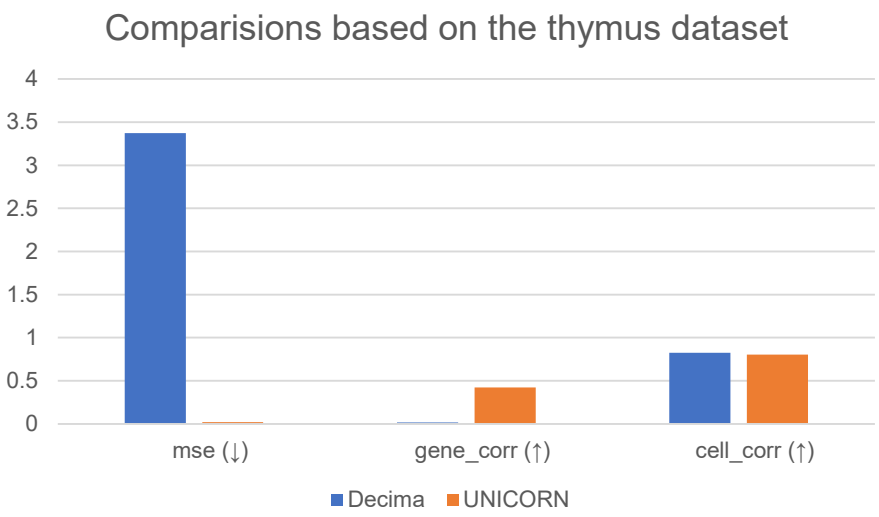
F Can we use sequence-to-function models to validate variant effects in cell-type levels?

Here, we collected the fine-mapping variants from CausalDB [70] and utilized the effective alleles and non-effective alleles to generate different groups of prediction results for cells, which correspond to reference output and variant output. We ranked the p-values of different fine-mapping variants in ascending and selected the first 100 variants as well as the last 100 variants to validate the contributions of variants for gene expression levels. Our results are shown in Extended Data Figures 18 (a) and (b). Based on these two figures, we found that the correlation of gene expression levels between reference outputs and variant outputs is consistently high ($\geq$0.99) (while the MSE is consistently low ($\sim$1e-7)) for both groups. We further computed the Rank-sum test statistics for both the correlation coefficients and mean squared errors between reference outputs and variant outputs using these two groups, and both of them did not show significance (p-value=0.33 for correlation coefficients, and p-value=0.50 for mean squared errors). Therefore, we did not consider further analyses for the effect of a single variant on gene expression prediction. One possible reason to explain the performance is that the variants in the genome always work as a group [11, 71, 72]. Another reason is, the model Enformer was pre-trained based on the reference genome rather than the individual genome, which could also introduce a potential bias for individualized prediction.

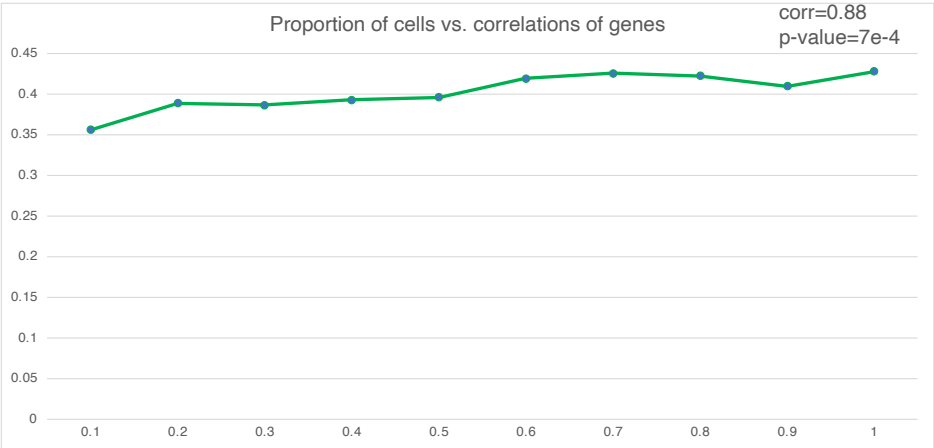
G Supplementary figures



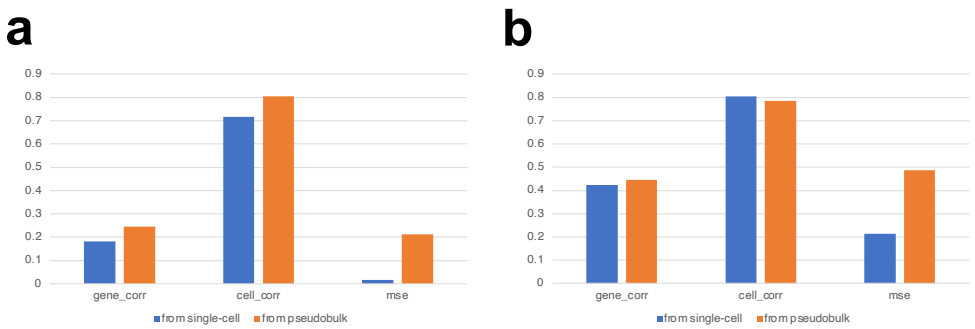
Extended Data Fig. 1 Comprehensive evaluations for gene expression predictions from sequences. The triangle shape represents the mean value, and the black dashed line represents the median value. We report the results based on box plots. (a) Results of cell-level correlations across different methods based on the thymus dataset. (b) Results of MSE across different methods based on the thymus dataset. (c) Results of cell-level correlations across different methods based on the PBMC dataset. (d) Results of MSE across different methods based on the PBMC dataset.



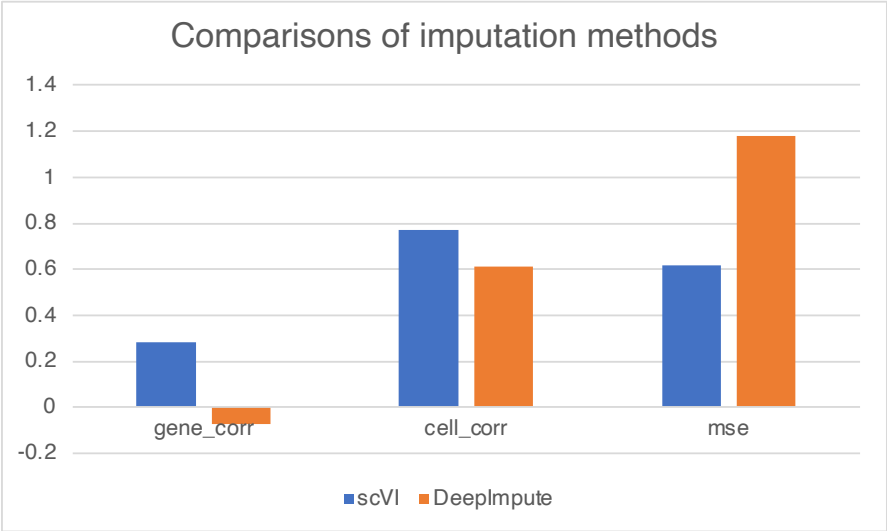
Extended Data Fig. 2 Comparison between UNICORN and Decima based on the thymus dataset. To make a fair benchmark, we select the overlapped cell types and tissue from Decima's prediction results.



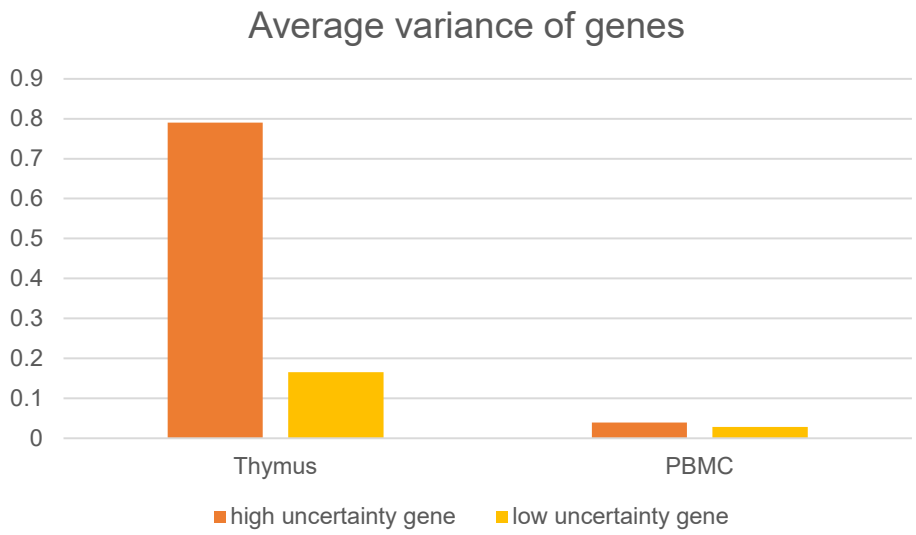
Extended Data Fig. 3 The relationship between the proportion of cells used for training and the correlation coefficients computed based on genes. The correlation is computed based on Pearson correlation.



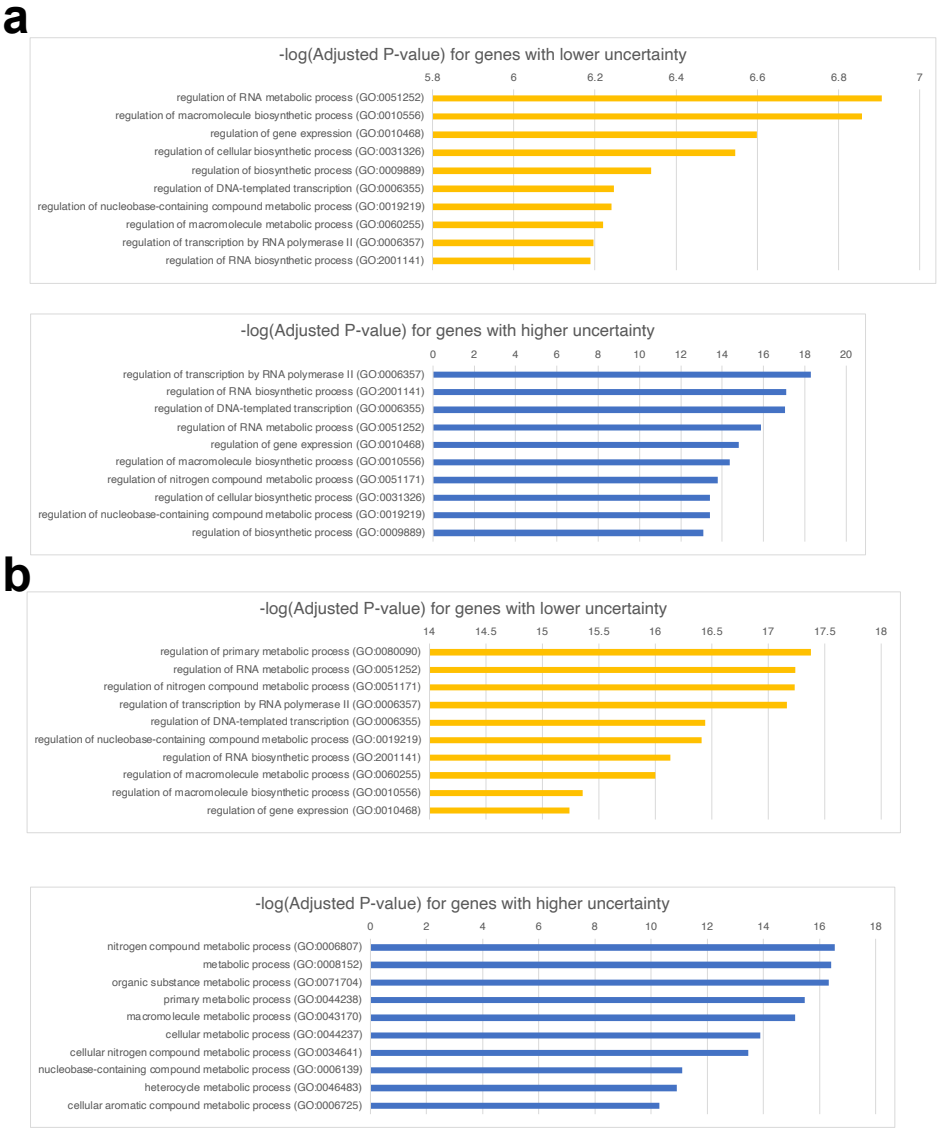
Extended Data Fig. 4 Comparisons between gene expression prediction based on single-cell data as input and gene expression prediction based on pseudo-bulk data as input. Both results are evaluated based on the pseudo-bulk level. (a): Comparisons of different metrics between two settings based on the thymus dataset. (b): Comparisons of different metrics between two settings based on the PBMC dataset.



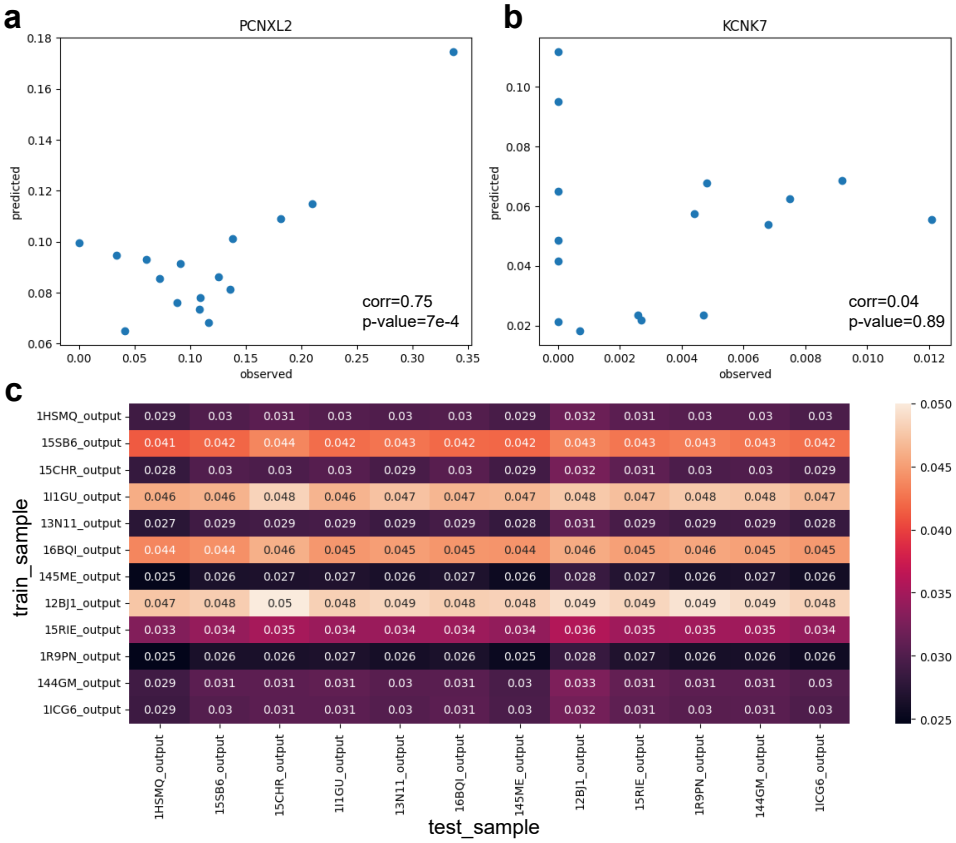
Extended Data Fig. 5 MSE under different imputation methods.



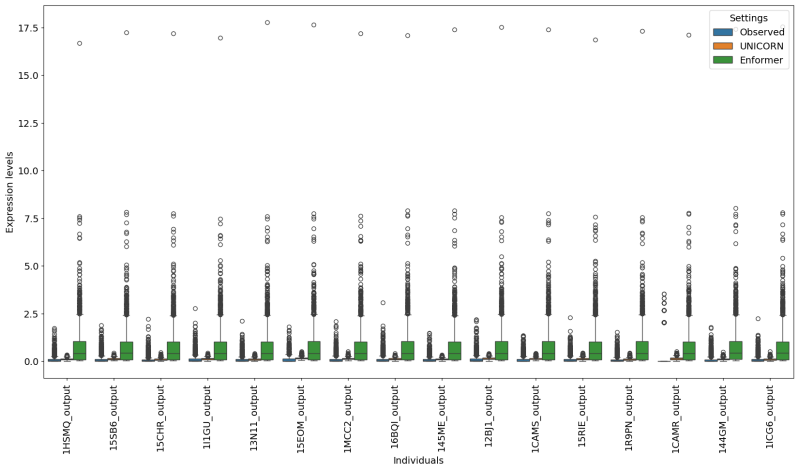
Extended Data Fig. 6 Average variance of genes for the two different datasets.



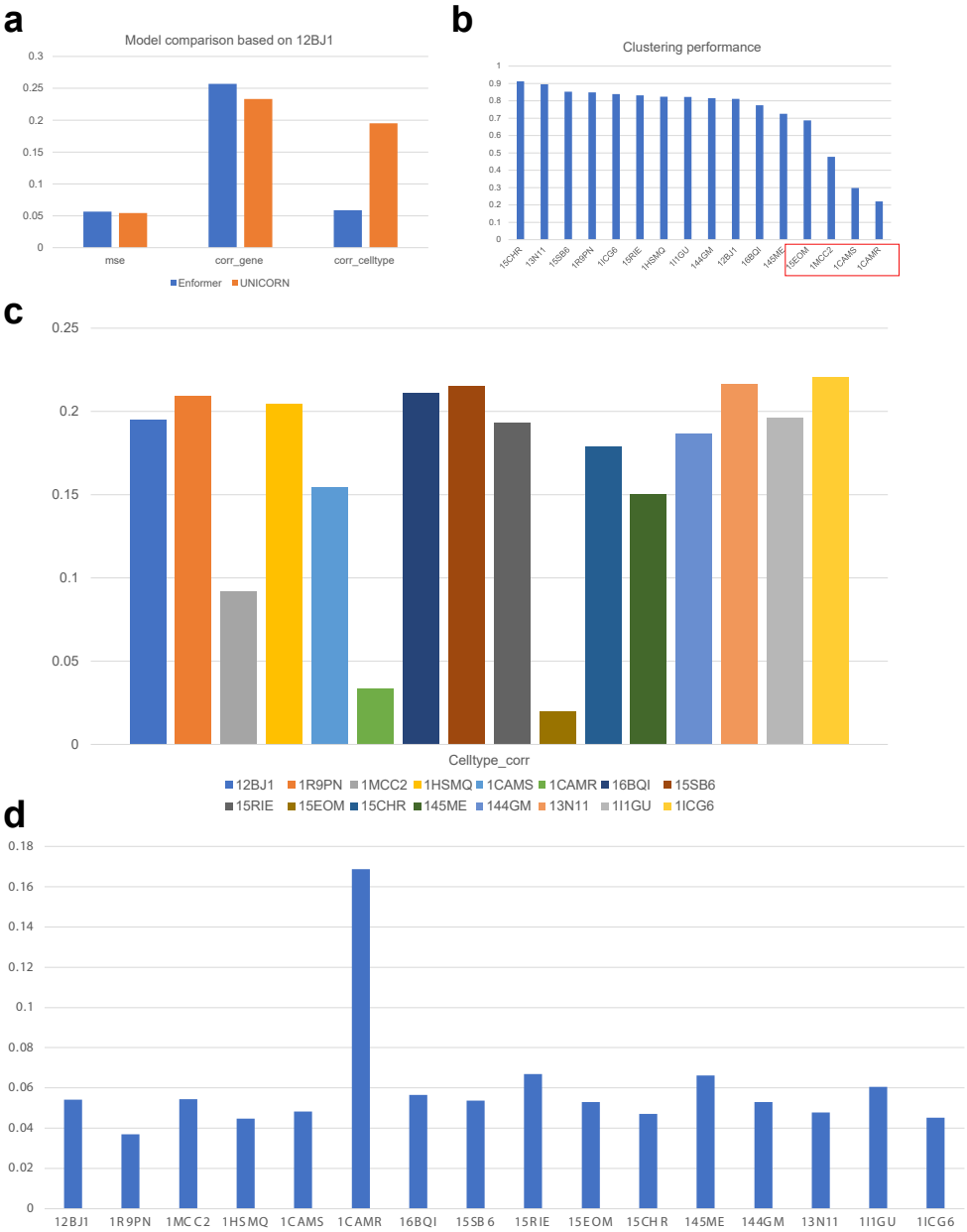
Extended Data Fig. 7 GO Enrichment Analysis for genes with different uncertainty levels. (a) Pathways from genes with lower uncertainty and higher uncertainty for the thymus dataset. (b) Pathways from genes with lower uncertainty and higher uncertainty for the PBMC dataset.



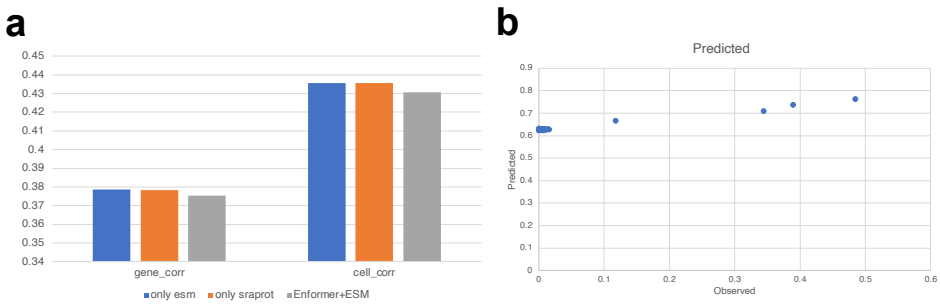
Extended Data Fig. 8 Analysis of gene expression predictions for individualized scRNA-seq datasets. (a) The performance comparison between Enformer and UNICORN based on the sample 12BJ1. (b) The average clustering scores across different individuals based on the corresponding scRNA-seq datasets. (c) The results of correlation coefficients computed based on cell types across different samples. (d) The results of MSE computed based on genes and cells across different samples.



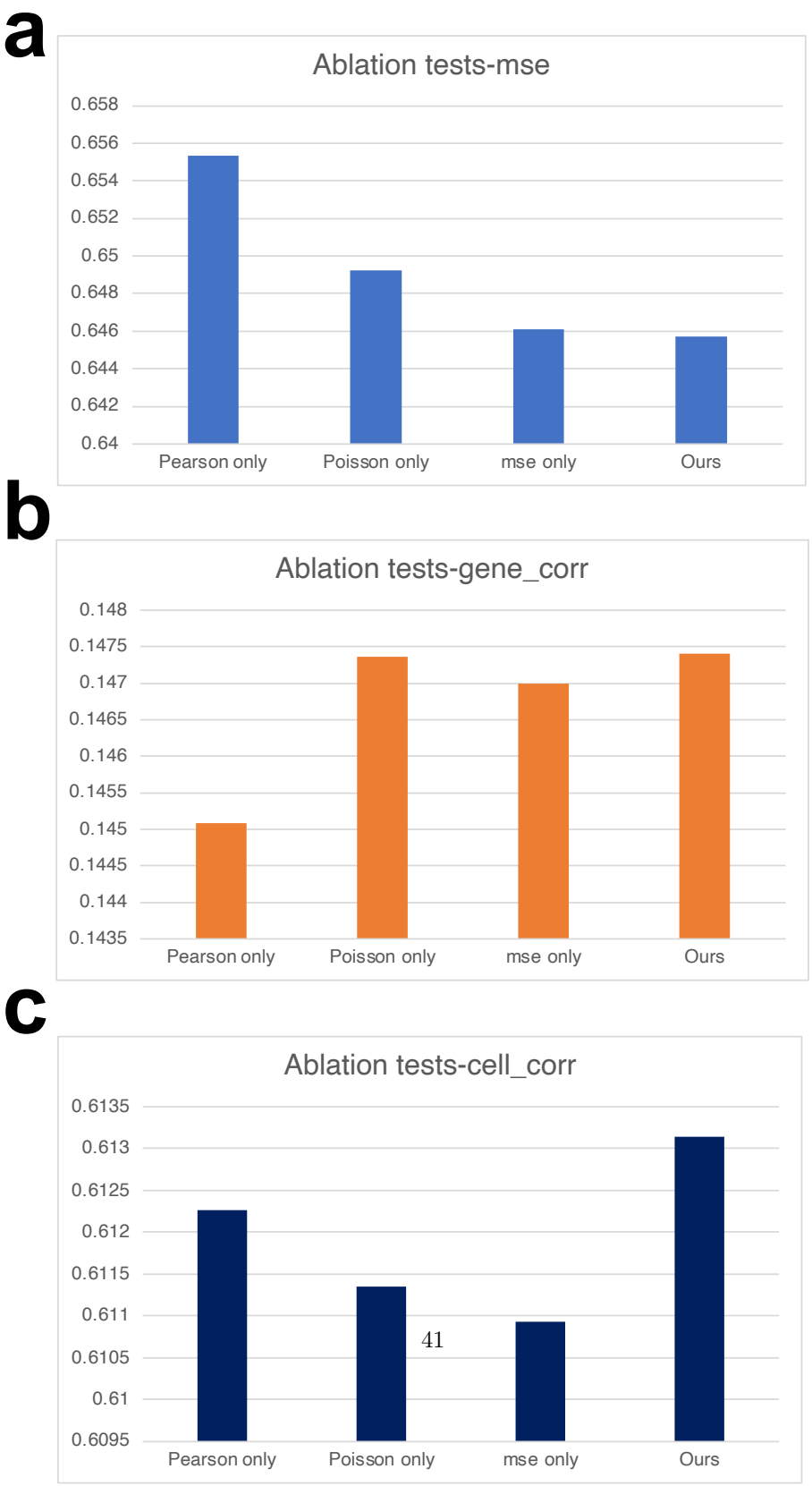
Extended Data Fig. 9 Comparisons of gene expression levels across individuals. We assign different colors for the observed values and predicted values.



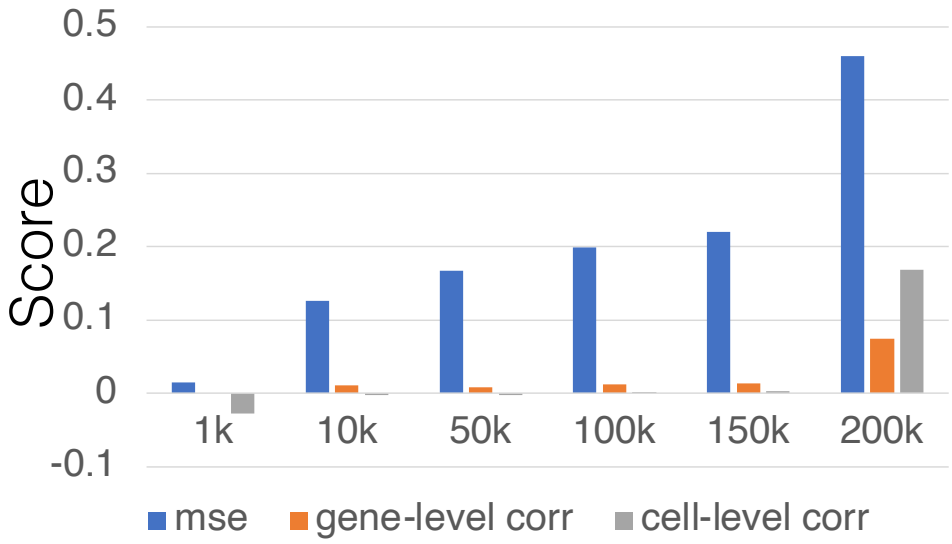
Extended Data Fig. 10 Examples of genes with different prediction performances and the results of cross-individual prediction. (a) The correlation between observed PCNXL2 expression and predicted PCNXL2 expression (good example). (b) The correlation between observed KCNK7 expression and predicted KCNK7 expression (poor example). (c) The MSE value of cross-individual prediction result based on UNICORN. The rows representing training samples and the columns represent the testing samples.



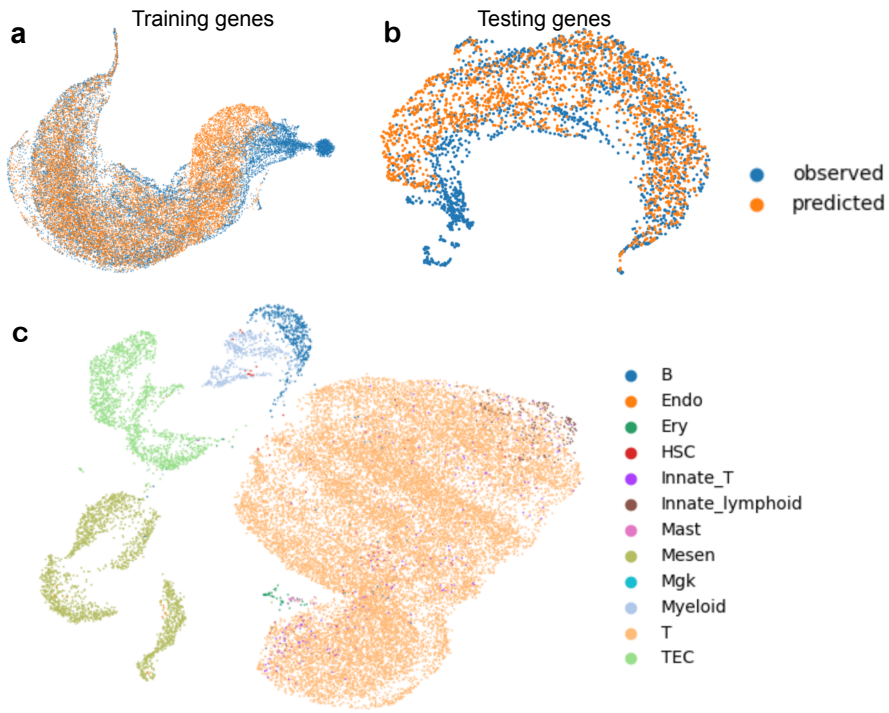
Extended Data Fig. 11 Analysis the factors in affecting multi-omic prediction. (a) The ablation test for different approaches of embedding combination. Here ESM2 [60] and SaProt [73] are two models used to generate protein embeddings. (b) The scatter plot for visualizing the relationship between observed peak information and predicted peak information for different cell types measured in 10X Multi-Omic dataset.



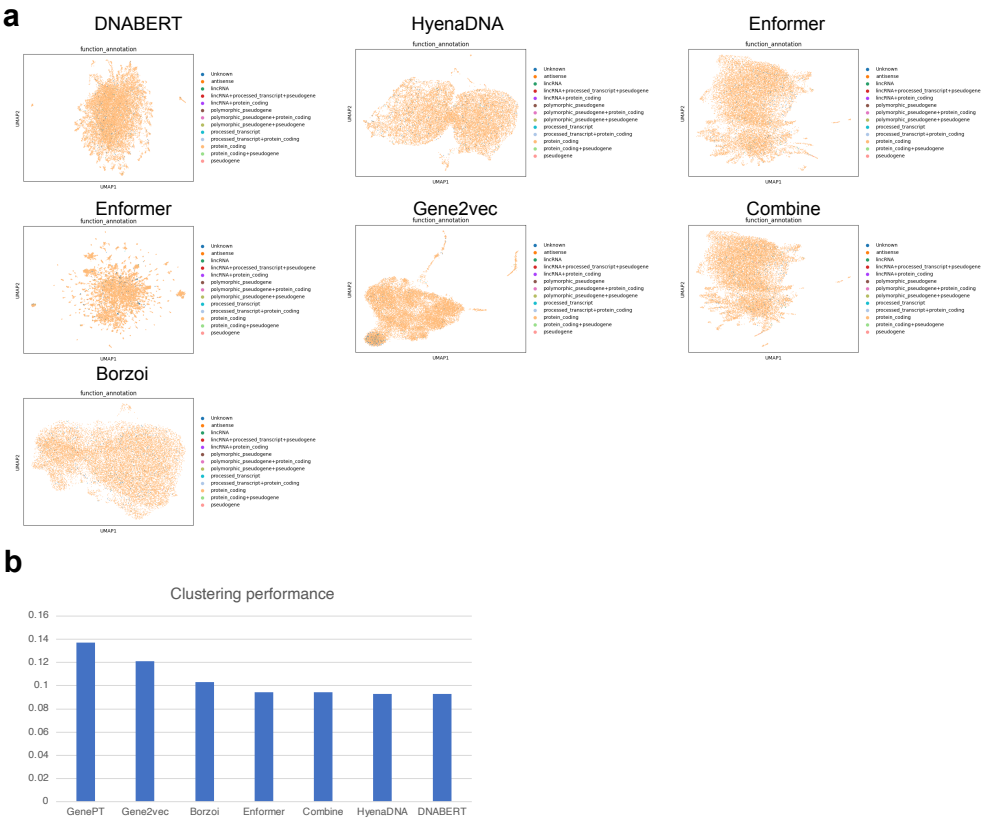
Extended Data Fig. 12 Results of ablation tests. We report the averaged score for each metric. (a) The correlation coefficients of genes under different choices of loss function components. (b) The correlation coefficients of cells under different choices of loss function components. (c) The MSE is under different choices of loss function components



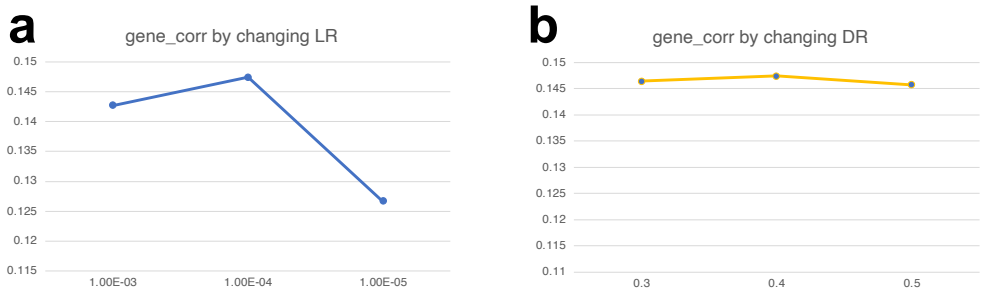
Extended Data Fig. 13 Comparisons for single-cell gene expression prediction under different context lengths. We visualize three metrics (MSE, gene-level correlation score, and cell-level correlation score) by showing their averaged values under different context lengths from 1k to 200k.



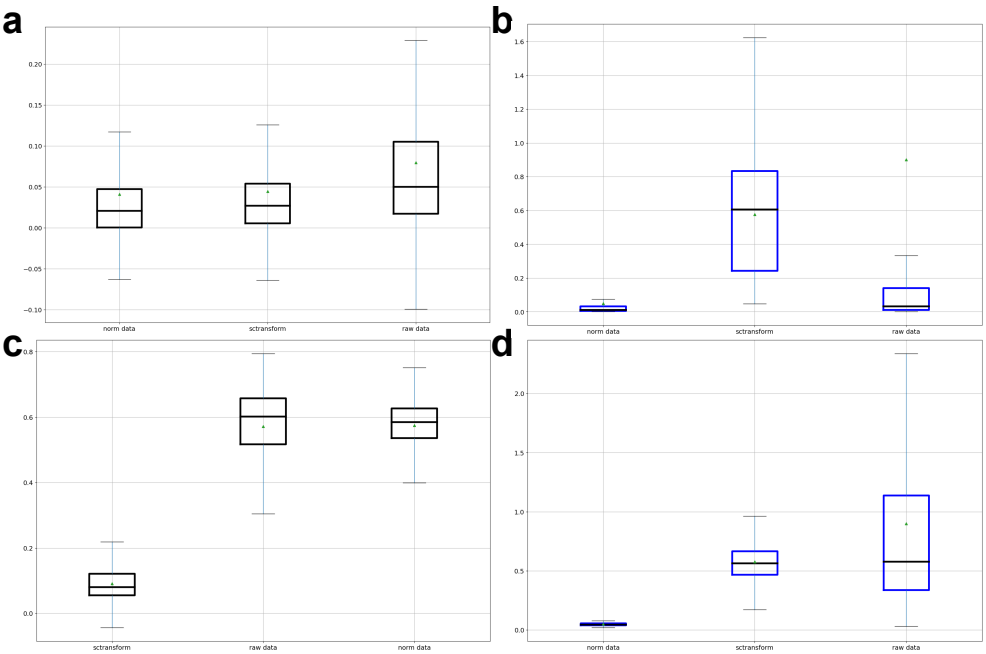
Extended Data Fig. 14 The visualization results of atlas-level data prediction. (a) The UMAP plots for training gene embeddings from observed gene expressions and predicted gene expressions. (b) The UMAP plots for testing gene embeddings from observed gene expressions and predicted gene expressions. (c) The UMAP plots for cells with predicted expression levels from the thymus atlas dataset, colored by cell types.



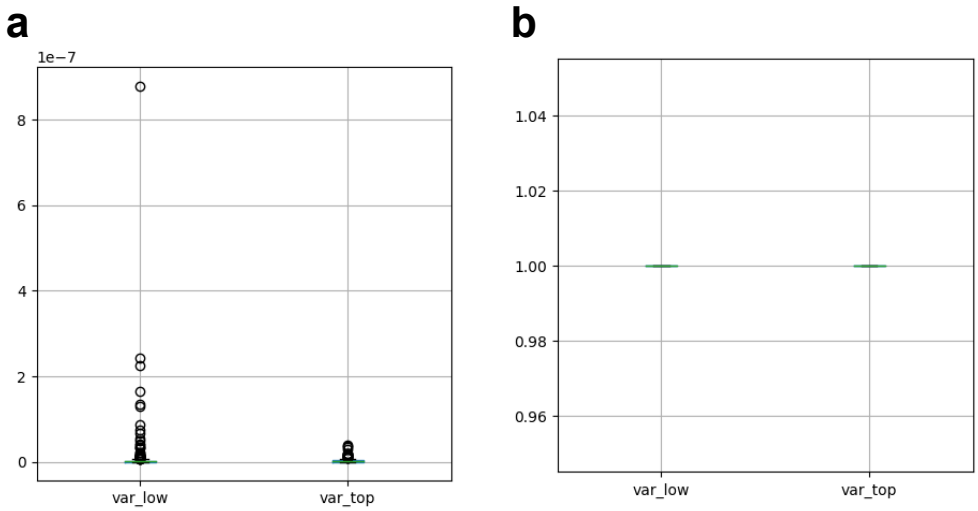
Extended Data Fig. 15 Performances of clustering based on different gene embeddings. (a) The UMAP plots of different gene embeddings colored by functional annotations. (b) The averaged scores of different gene embeddings.



Extended Data Fig. 16 Model performances under different settings of hyper-parameters. (a) The effect of changing LR. (b) The effect of changing DR.



Extended Data Fig. 17 Comparisons between raw data mode (raw data), log-normalized data mode (norm data), and sctransform data mode (sctransform). 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 PBMC dataset. (b) Results of gene-level MSE across different methods based on the PBMC dataset. (c) Results of cell-level correlations across different methods based on the PBMC dataset. (d) Results of cell-level MSE across different methods based on the PBMC dataset.



Extended Data Fig. 18 The difference of metrics computed based on effective allele and non-effective allele. Here the legend *var_low* represent last 100 alleles and the legend *var_top* represents the top 100 alleles. The gene expression is computed at the pseudo-bulk level. (a) The distribution of MSE between the predicted gene expression and the observed gene expression. (b) The distribution of gene-level correlation coefficients between the predicted gene expression and the observed gene expression.