Transcriptome Complexity Disentangled: A Regulatory Molecules Approach

Amir Asiaee [*†], Zachary B. Abrams [‡], Heather H. Pua [§], and Kevin R. Coombes [¶]

Abstract

Transcription factors (TFs) and microRNAs (miRNAs) are fundamental regulators of gene expression, cell state, and biological processes. This study investigated whether a small subset of TFs and miRNAs could accurately predict genome-wide gene expression. We analyzed 8895 samples across 31 cancer types from The Cancer Genome Atlas and identified 28 miRNA and 28 TF clusters using unsupervised learning. Medoids of these clusters could differentiate tissues of origin with 92.8% accuracy, demonstrating their biological relevance. We developed Tissue-Agnostic and Tissue-Aware models to predict 20,000 gene expressions using the 56 selected medoid miRNAs and TFs. The Tissue-Aware model attained an R^2 of 0.70 by incorporating tissue-specific information. Despite measuring only 1/400th of the transcriptome, the prediction accuracy was comparable to that achieved by the 1000 landmark genes. This suggests the transcriptome has an intrinsically low-dimensional structure that can be captured by a few regulatory molecules. Our approach could enable cheaper transcriptome assays and analysis of low-quality samples. It also provides insights into genes that are heavily regulated by miRNAs/TFs versus alternative mechanisms. However, model transportability was impacted by dataset discrepancies, especially in miRNA distribution. Overall, this study demonstrates the potential of a biology-guided approach for robust transcriptome representation.

1 Introduction

The transcriptome comprises the entire set of messenger RNA molecules within an organism or cell and its regulation controls gene expression. Due to its ease of measurement, the transcriptome has played a significant role in advancing our understanding of differentiation [1], development [2], and disease [3]. The transcriptome is both an essential contributor to the cell's underlying biology and a quantitative measure that provides a mathematical and concrete representation of the cell state [4, 5].

Although RNA-Seq has become the most popular gene expression measurement technology due to the significant drop in its cost since 2008 [6], it is not always effective or feasible for measuring expression in biological or clinical specimens. For instance, when measuring gene expression on a scale of millions of experiments to determine cellular function or response (such as response to drugs or gene perturbations), the low cost per sample (ranging from $10 to $200, depending on the technology) might still be prohibitively expensive [7]. Additionally, when high-quality RNA samples are unavailable, alternative sensitive techniques like qPCR may be preferable, although they are not ideal for whole transcriptome analysis due to the need for specific primers,

[*]Department of Biostatistics, Vanderbilt University Medical Center, 2525 West End Avenue, Nashville, TN 37203, USA

[†]To whom correspondence should be addressed. Tel: +1 615 3222067; Email: amir.asiaeetaheri@vumc.org

[‡]Institute for Informatics, Washington University, 4444 Forest Park Ave, St. Louis, MO 63108, USA

[§]Department of Pathology, Microbiology, and Immunology, Vanderbilt University Medical Center, 1161 Medical Center Dr, Nashville, TN 37240, USA

[¶]Department of Population Health Science, Medical College of Georgia, 1120 15th St, Augusta, GA 30912, USA

limited throughput, and technical variability [8].

To address these challenges, the NIH LINCS project [7] developed a new, low-cost, high-throughput method for measuring a "reduced representation" of the gene expression profile consisting of 978 landmark genes and 80 constant control genes that together are called L1000 assay. This method enables the inference of $11,350$ non-measured transcripts with 81% accuracy as measured by the correlation score. The authors employed this approach to scale the Connectivity Map project's samples from 492 original samples [9] to more than 1.3 million samples [7], enabling the discovery of functional connections among diseases, genetic perturbations, and drug actions. In the case of low-quality RNA samples, the same set of predictive models can be theoretically used for predicting transcriptome from the qPCR measurement of the L1000 genes.

Notably, the L1000 genes are chosen based on a pure data-driven approach rather than utilizing prior biological knowledge, as highlighted by Subramanian et al. [7]. In short, their approach is based on clustering of genes in the Principal Component space and selecting the centroids by an iterative peel-off procedure. Surprisingly, the authors reported that the landmark genes were not enriched in any specific functional class, such as transcription factors, as determined by hypergeometric overlap statistics [10].

The pure data driven selection method of predictive genes overlooks the causal relationships within the transcriptome, as some parts of the transcriptome act as the root causes or master regulators of others. For instance, certain transcripts produce transcription factor proteins that regulate the transcription of other genes [11]. Therefore the abundance of former transcripts can be considered the causes of the latter ones. We hypothesize that taking into account the causal relationships between transcriptome components when making predictions can have two significant implications: firstly, by treating the causes as the main predictors for the entire transcriptome, the size of the "reduced representation" of the whole transcriptome could be substantially decreased; secondly, incorporating causal predictors into the model could result in more robust predictions across various environments.

In this study, we use Transcription Factor (TF) and microRNA (miRNA), two classes of regulatory molecules known to causally regulate transcription, as predictors to demonstrate that by measuring only 28 of each class, we can explain 70% variation of the transcriptome, i.e., expression of approximately 20,000 genes. Our findings suggest that the transcriptome has a low-dimensional structure that can be captured by measuring the expression of only 56 regulatory molecules. Each of our selected regulatory molecules represents a cluster of highly co-expressed regulatory molecules, which we term as a regulatory component. Through gene enrichment analysis, we show that regulatory components consist of genes that are co-expressed and work together to control various biological processes [12, 13], which is in contrast to the selected L1000 landmark genes. Additionally, we show that our approach requires two orders of magnitude fewer measurements than the traditional L1000 approach while achieving comparable whole transcriptome prediction accuracy. This could lead to the development of cheaper transcriptome assays and accommodate low quality RNA sample scenarios. Furthermore, we provide biological insights into potential reasons for certain genes being easy or hard to predict using TFs and miRNAs by analyzing the set of genes that are poorly or well-explained by our model. Lastly, we evaluate the transportability of our model by assessing its performance in whole transcriptome prediction on three previously unseen datasets.

2 MATERIALS AND METHODS

2.1 Data Description

2.1.1 Main Dataset.

The main data used in this study were collected from The Cancer Genome Atlas (TCGA) [14], a public data repository that contains clinical and omics data for over 10,000 patient samples from 32 differ-

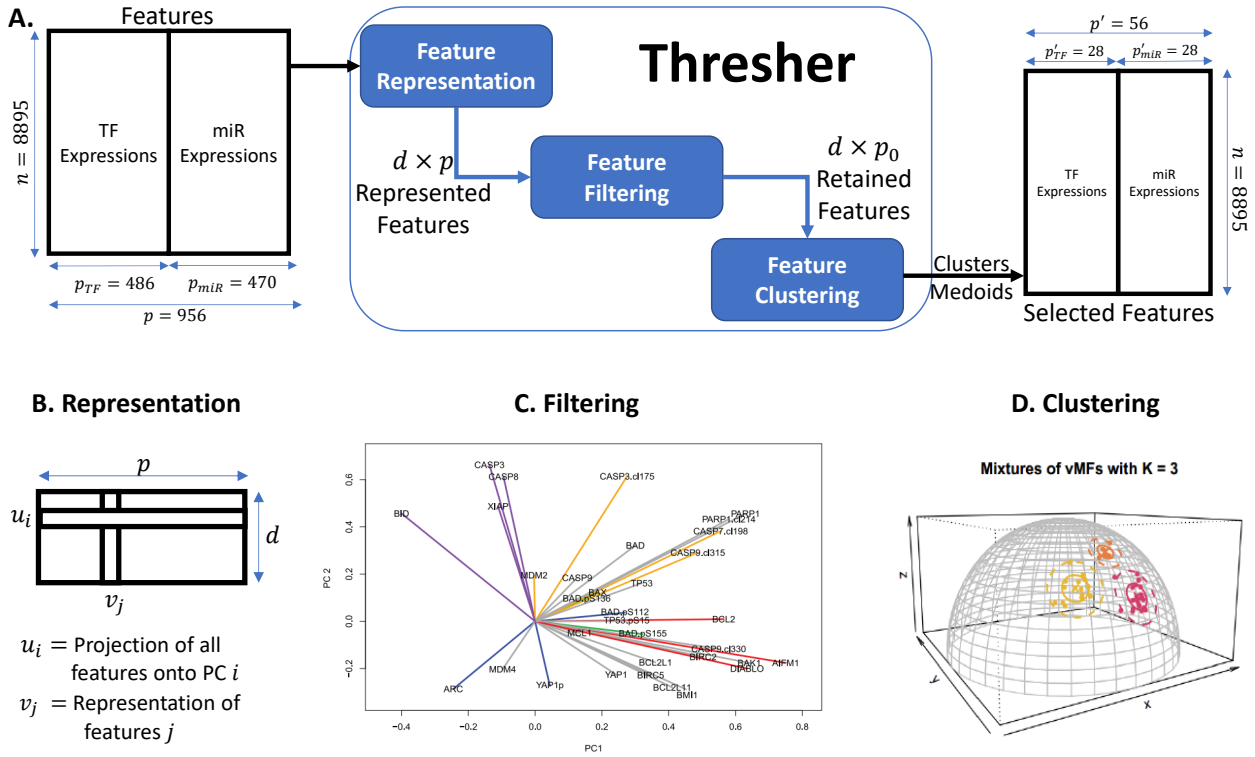

Figure 1: **Overview of the Feature Selection Pipeline.** **(A)** Overview of the Thresher algorithm, detailing its input, internal processes, and output. The regulatory molecules of the TCGA dataset, i.e., transcription factors and microRNAs serve as the input. Thresher first learns a representation of each TF and microRNA, then filters out the less significant ones, and finally clusters the remaining miRs and TFs. A medoid from each cluster is chosen to represent it. The resulting selected regulatory molecules are two orders of magnitude fewer than the initial set but retain comparable biological predictive power. Panels **(B)**, **(C)**, and **(D)** provide detailed illustrations of each internal step of the Thresher. **(B) Feature Representation.** Features are represented using PCA applied to the feature-by-sample matrix. Each feature i is represented by $\nu_i \in \mathbb{R}^d$, where d is determined through Bayesian model selection. **(C) Feature Filtering.** Feature representations are visualized using only their first two principal components, revealing significant variance in the lengths of loadings. Features with smaller loadings are filtered out; specifically, a feature is discarded if $\|\nu_i\|_2 < 0.35$. **(D) Feature Clustering.** Post-filtering, the lengths of feature representations are considered less critical. The clustering then focuses on the directions of the features, $\nu_i/\|\nu_i\|_2$, which requires clustering of points on a hypersphere, where features are assumed to have mixtures of von Mises-Fisher distributions.

ent cancer types. To obtain the transcript data, we used the FireBrowse web portal and selected samples with matched mRNA and microRNA sequencing data. We obtained $n = 8895$ samples from 31 different cancer types. Glioblastoma, despite having a substantial number of samples, was excluded as almost all of its samples lacked the necessary microRNA measurements.

We downloaded the mRNA expression level data that were quantified using the RSEM method [15] and normalized using FPKM-UQ (Upper quartile normalized fragments per Kilobase of transcript per Million mapped reads). We obtained the microRNA expression data in the form of counts of mature microRNA, normalized to reads per million mapped reads (RPM). We log2 transformed both the mRNA and miRNA datasets. We filtered the data by removing miRNAs with zero read counts in 75% of the samples, resulting in p =470 miRNAs remaining [13]. Using the Transcription Factor Catalog [16], we identified 486 human transcription factors and separated them from the 20289 measured gene expressions [12]. The resulting dataset consisted of $8895 \times (486 + 470)$ dimensions, comprising TFs and miRNAs as our original features and the rest of the transcriptome with 19803 gene expressions for the 8895 samples as outcomes.

The specific cancer types from TCGA that we used, along with their respective abbreviations, are presented in the Supplemental Data. It is important to understand that certain TCGA cancer types, such as Lung Adenocarcinoma and Lung Squamous Cell Carcinoma, originate from the same tissue. They essentially represent different subtypes of a primary cancer based on that tissue's lineage. When training our model on TCGA data and considering the cancer type labels, a more precise name for our model would be "Cancer Tissue-Aware." However, due to the broader applicability of our model beyond just TCGA and for the sake of simplified terminology, we will use "Tissue-Aware" as the descriptor for our model.

2.1.2 Auxiliary Datasets

. To ascertain the influence of normalization on our results, in addition to FPKM-UQ normalized data, we conducted all experiments using a TPM (Transcription Per Million) normalized version of the TCGA data. Additionally, to evaluate the transportability of our whole transcriptome prediction model trained on TCGA, we assessed its performance on three distinct datasets: GTEx (Genotype–Tissue Expression [17]), TARGET (Therapeutically Applicable Research to Generate Effective Treatments [18]), and CCLE (Cancer Cell Line Encyclopedia [19]). These datasets were chosen due to their broad tissue and cell type coverage.

In detail, CCLE offers an exhaustive genetic analysis of over 1,000 human cancer cell lines, making it an invaluable repository for advancing new cancer treatments and probing the intricate mechanisms driving tumor development. GTEx, on the other hand, sourced its samples from 54 non-diseased tissue sites spanning nearly 1,000 individuals, providing in-depth perspectives on gene expression dynamics and regulatory patterns across various human tissues. Lastly, TARGET is dedicated to understanding the molecular alterations propelling five pediatric cancers: ALL, AML, kidney cancer, neuroblastoma, and osteosarcoma.

While all three datasets encompass RNA-Seq data, only the CCLE offers distinct measurements for microRNA. For our transportability experiment, we retrieved log2(TPM + 0.001) RNA-Seq data for TCGA, GTEx, and TARGET from the Xena browser [20], and both log2(TPM + 0.001) RNA-Seq data and RPM normalized microRNA data for CCLE from the DepMap project [21]. The Xena browser, by providing uniformly realigned and re-called gene and transcript expression data for TCGA, TARGET, and GTEx, assists in minimizing variances in gene-level RNA-Seq data attributable to different bioinformatics pipelines.

Our study focused on 942 samples from CCLE, which contained measurements for 18,706 mRNAs and 743 microRNAs. The count of RNA-Seq samples from the Xena Browser's TCGA, GTEx, and TARGET datasets stands at 10,013, 7,792, and 734, re-

spectively, with all three datasets capturing data for 19,069 genes.

2.2 Feature Selection

In this study, we aim to understand the computational complexity of transcriptomes by selecting a low-dimensional representation that captures most of the information about the sample's state. Specifically, we focus on regulatory molecules miRNAs and TFs as features and aimed to identify a small subset of them that can be used in downstream predictive analyses while maintaining comparable performance to using the whole transcriptome.

We chose feature selection instead of dimensionality reduction methods like Principal Component Analysis (PCA) because measuring the corresponding low dimensional latent features, i.e., linear combinations of all TFs and miRNAs, requires measuring the all regulatory molecules, which can be expensive and challenging for low-quality RNA samples. Additionally, interpreting the biological relevance of latent features that are superpositions of numerous genes can be challenging [22]. We chose not to use prediction-guided feature selection methods [23], such as LASSO [24], as our goal was to choose features in a task-agnostic manner, ensuring that the performance of any downstream prediction task remains comparable to the situation where the entire transcriptome was measured. For example, employing LASSO for whole transcriptome prediction could yield varying sets of predictive miRNAs and TFs for each gene, making the measurement of nearly all miRNAs and TFs a necessity. This goes against our aim of showing that a subset of regulatory molecules can efficiently represent the transcriptome.

Our feature selection process employs the Thresher package [25], which automatically selects the optimal number of clusters for clustering TFs and miRNAs into regulatory components, Figure 1. We then select the medoid of each cluster as the representative gene for the corresponding regulatory component. The Thresher pipeline begins by applying PCA to the feature-by-sample matrix (instead of the sample-by-feature matrix) to derive a low-dimensional representation of each feature. To identify the number of significant principal components, we use a variant of the Bayesian model selection method [26], implemented in the PCDimension package. We further refine the feature set by removing outliers using a length threshold for each feature representation. A threshold of 0.35 was chosen based on several simulation studies [25]. Finally, we cluster the *directions* of feature representation vectors, corresponding to points on the hypersphere, by von Mises-Fisher mixture model [27]. We select the optimal number of regulatory molecule clusters using the Bayesian Information Criterion [28]. To demonstrate that the selected regulatory molecules capture most of the transcriptome's information, we perform various predictive tasks summarized in the subsequent sections.

2.3 Assessing Predictive Power of Selected Regulatory Molecules

2.3.1 Tissue Classification Using Selected Regulatory Molecules.

To show that the selected 56 regulatory molecules are biologically informative, we use them to predict the cancer type of TCGA samples. This is a multi-class classification problem [29] with 31 outcome labels. To train our classifier, we use a Support Vector Machine (SVM) algorithm [30] and report the average generalization error on hold-out data through 10-fold cross-validation.

2.3.2 Transcriptome Prediction Using Selected Features and Tissue Type Information.

Our goal was to test whether the 56 selected features provide an accurate representation of the entire transcriptome. In the first experiment, we use linear regression to predict the entire transcriptome using the selected features. We have approximately 20,000 regression tasks, with each task consisting of 56 features and ∼9,000 samples. For each outcome gene, we obtain a model with 57 parameters, consisting of 56 coefficients and one intercept. Since we do not use cancer tissue type information in prediction, we call this model **Tissue-Agnostic**.

In the second experiment, we use cancer tissue type information available in the TCGA dataset to

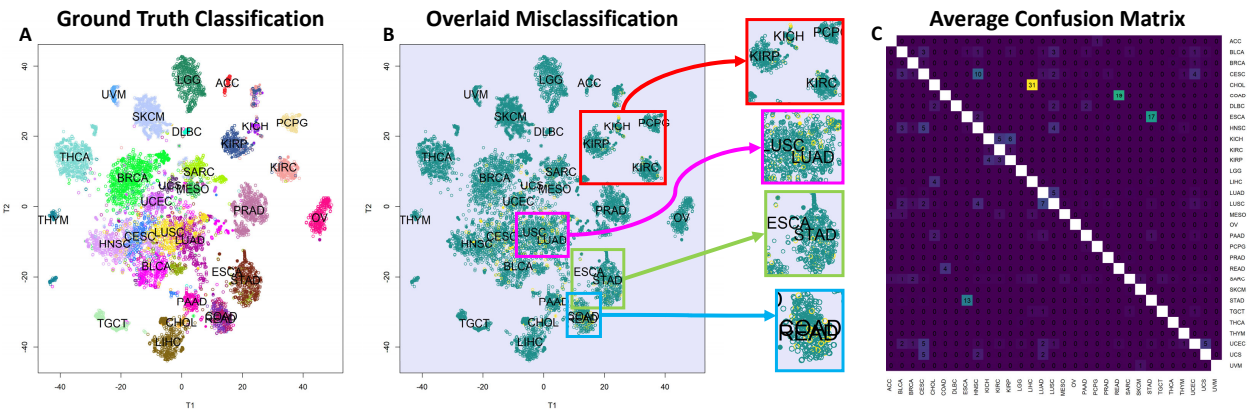

Figure 2: **Cancer Type Classification and Visualization Through Regulatory Molecules. (A)** A 2D t-SNE plot of the 8895x56 input TF and miRNA expression matrix, consisting of selected TFs and miRNAs, with colors corresponding to known cancer types, as listed in Supplemental Data, showcasing the ability to differentiate most cancers using 56 regulatory molecules. **(B)** An enlarged t-SNE plot focusing on misclassified samples, marked in yellow, indicates that classification errors are more common in highly similar cancer types. **(C)** An average confusion matrix generated from 10-fold cross-validation, exposing the error pattern of the SVM classifier with 92.8% accuracy. The numbers within each cell of the heatmap represent the average percentage of times, across 10-fold cross-validation, that samples from one class (row) were predicted to belong to another class (column). A value of 100% would indicate that all samples of a given class were consistently misclassified as another specific class across all cross-validation folds. Diagonal cells representing correct classifications have been left blank for clarity. Lighter matrix entries (excluding diagonals) represent larger errors, and the error pattern predominantly coincides with the t-SNE visualization of errors in panel B. Classifying similar cancer pairs like STAD and ESCA, COAD and READ, LUSC and LUAD, higher-order similar cancers such as KIRP, KIRC, and KICH, as well as squamous cancers (or cancers with a squamous cell subtype) including LUSC, CESC, HNSC, and BLCA, is challenging.

improve our model and call it **Tissue-Aware** model. Instead of predicting the outcome for each gene as a single task, we predict the outcome in each tissue as a separate task, resulting in 31 tasks for TCGA. We first fit a global regression model across all tissues and then fit the residuals of the global model in each tissue through a secondary regression. This can be considered a form of multi-task learning [31], where a shared intercept is learned for all tasks along with one linear model per task [32, 33, 13].

Mathematically, to predict the expression of gene g, denoted as $\mathbf{y}_t^g \in \mathbb{R}^n$, in tissue t using the selected expression matrix $\mathbf{X}_t$ of TFs and miRs, we determine a coefficient vector $\beta_t^g \in \mathbb{R}^{56}$ and an intercept α_t specific to each tissue. Additionally, we introduce a global (cross-tissue) intercept α_0. This Tissue-Aware optimization is expressed as:

$$(\alpha_0, \forall t : \alpha_t^g, \beta_t^g) = \underset{\alpha_0^g, \forall t: \alpha_t^g, \beta_t^g}{\arg\min} \sum_t \|\mathbf{y}_t^g - \mathbf{X}_t \beta_t^g - \alpha_t^g - \alpha_0^g\|_2^2$$

2.4 Biological Insights from Evaluating Transcriptome Prediction Accuracy

To gain insights into the biological implications of transcriptome prediction accuracy, we evaluated the Tissue-Agnostic and Tissue-Aware models and examined how much variance in the transcriptome they

accounted for. Our analysis focused on the distribution of R^2 values, which represents the proportion of variation in gene expression that each model explains. We created a histogram of R^2 values for the nearly 20,000 genes predicted and fitted a beta mixture model to identify poorly, moderately, and well-explained genes from the histogram. We applied the expectation-maximization algorithm with a range of mixture components ($k = 1, 2, 3, 4$) and used the Bayesian Information Criterion (BIC) score to determine the optimal number of components. The R^2 distribution of both models showed evidence of two beta distribution components, which we labeled as **poorly and well-explained components**, Figure 3.

To gain insight into which biological pathways are influenced by TFs and miRs, and which pathways are less affected by these regulatory molecules, we performed gene enrichment analysis using the ToppGene platform [34]. Specifically, we focused on genes in the bottom 5th percentile of the poorly explained component and in the top 99th percentile of the well-explained component for each model. Our goal was to identify pathways that are heavily controlled by these regulatory molecules, regardless of tissue involvement in the modeling. To accomplish this, we examined the intersection of pathways enriched for poorly-explained genes in each model, as well as for well-explained genes.

2.5 Evaluation of Normalization Method Influence on Results

Throughout our study, we have primarily utilized the FPKM-UQ normalization method for TCGA's RNA-Seq data. To ensure our results are not solely dependent on this specific normalization method, we conducted all previous experiments—including tissue classification, transcriptome prediction, and gene enrichment analysis—using TPM-normalized TCGA data. TPM is recognized for better preservation of biological signals compared to other normalization techniques [35].

2.6 Comparison of Performance between Selected Regulatory Molecules and L1000 Genes

The L1000 Gene Expression Signature Profiling Technology and the performance of the 978 landmark genes it uses for whole transcriptome prediction have been widely studied for identifying new therapeutic targets and drug combinations [7]. However, due to the lack of access to the data set used for their gene selection and prediction, we could not test the performance of our 56 selected regulatory molecules on the same data set. To ensure a fair comparison, we instead used the L1000 genes to fit TCGA data from scratch.

We compared our Tissue-Aware model, which employs 56 genes, and the linear model, which uses L1000 genes. To evaluate the performance of the models, we computed their average R^2 scores and correlation scores. The correlation score, introduced in the L1000 landmark paper [7], measures the algorithm's overall performance in predicting gene expression as the fraction of genes that are "best inferred." A gene g is best inferred if the correlation $R_{gg} = \text{cor}(\hat{\mathbf{y}}^g, \mathbf{y}^g)$ is significant. To determine if R_{gg} is significant, a null distribution of correlation is made by randomly pairing genes g_1 and g_2 and computing $R_{g_1 g_2} = \text{cor}(\hat{\mathbf{y}}^{g_1}, \mathbf{y}^{g_2})$. The inference is deemed accurate if R_{gg} is higher than the 95th percentile of the null distribution. The linear model trained on the data from [7] that uses L1000 landmark genes as features has a reported correlation score of 81%.

2.7 Assessing the Model's Transportability Across Datasets

To validate the transportability of our Tissue-Aware transcriptome prediction model, we applied it to an independent dataset: the CCLE (Cancer Cell Line Encyclopedia) project [19]. The predictors for mRNA expression were sourced from the TFs and miRNAs selected by the Thresher using TCGA data, as shown in Figure 1.

However, integrating the selected miRNAs from TCGA with CCLE posed several challenges. Notably, TCGA includes measurements for both mature

miRNA directions, such as hsa-miR-145-3p and hsa-miR-145-5p, whereas CCLE only reports hsa-miR-145. Given the likely disparities in the abundance and targets of the 3p and 5p mature miRNAs, accurate matching to CCLE's miRNAs is crucial. Yet, it remains unclear if CCLE captures a dominant variant or an aggregate count of both. We opted to match the directional miRNAs from TCGA to their non-directional counterparts in CCLE, understanding that depending on CCLE's measurement methodology, this could significantly impact prediction performance. Furthermore, five miRNAs selected from TCGA were absent in CCLE. For these, we substituted the second "closest to the centroid" member from the respective miRNA clusters identified by Thresher. An additional concern arose from observing notable distributional differences for some matched miRNAs between TCGA and CCLE. We explore these topics further in subsequent sections.

With the Tissue-Aware model, our aim was to predict the transcription of 18,706 mRNAs across 942 CCLE samples. This task necessitated establishing a correspondence between the tissue (or cancer type) labels of TCGA and CCLE. Instead, we clustered TCGA samples with similar gene expressions into "Pseudo-Tissues." We employed the K-means algorithm [28], treated the number of clusters K as a hyper-parameter, and determined its optimal value via 5-fold cross-validation. To evaluate model performance on CCLE, each sample was mapped to a TCGA Pseudo-Tissue using the minimum Euclidean distance. The model trained on the associated Pseudo-Tissue was then deployed to predict the complete transcriptome of the given CCLE sample.

2.8 TF-Based Model Transportability Across Datasets

Considering the frequent omission of miRNAs in many large-scale RNA-seq datasets, we explored the feasibility of exclusively using TFs for whole transcriptome prediction. This shift towards TFs not only simplifies the prediction model but also widens the scope of applicable RNA-Seq datasets for our experiments. While we can potentially utilize any RNA-Seq dataset, we favored those with broad tissue representation, ensuring the robustness of our predictions across both cancerous and non-cancerous contexts.

With this objective in mind, we assessed the transportability of the Tissue-Aware model trained on TCGA using only TFs across three diverse datasets: GTEx [17], TARGET [18], and CCLE [19]. The learning procedure mirrored our previous approach: clustering TCGA samples into "Pseudo-Tissues" and training a Tissue-Aware model with TFs as input predictors. Following training, we leveraged this model to predict the transcriptomes of samples in the three aforementioned datasets, using their respective TFs as predictors.

3 RESULTS

3.1 Biological Relevance of Chosen Features

Our methodology identifies 28 miRNA clusters and 28 TF clusters, using their medoids as the selected features for our analysis. Prior studies have shown that the TFs [12] and miRNAs [13] clusters recognized by Thresher are associated with specific biological processes. Consequently, these clusters can be viewed as "regulatory components" representing cell states, as opposed to "principal components". We employed the unsupervised, nonlinear t-SNE projection technique [36] to the 8895×56 input matrix consisting of selected TFs and miRNAs and visualized the outcome in a two-dimensional plot shown in Figure 2A. To enhance our comprehension of the patterns, we color-coded the plot based on known cancer types. Overall, the majority of cancer types can be differentiated using the 56 chosen regulatory molecules. Utilizing the TPM normalized version of the TCGA data yielded comparable results. In this instance, we identified 28 miRNA clusters but observed a reduction in the number of TF clusters to 19. The corresponding t-SNE plot can be found in Supplement Figure S1A and B.

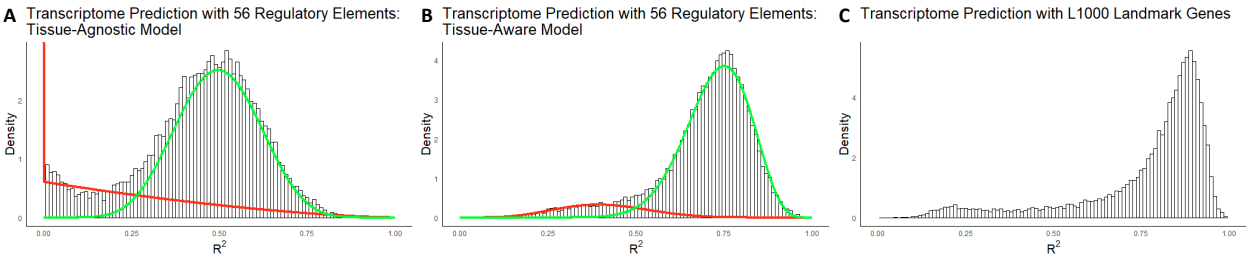

Figure 3: **Gene Expression Prediction and Comparison of Predictive Power between Selected Regulatory Molecules and L1000 Genes. (A)** and **(B)** R^2 histograms for predicting 20,289 gene expressions using 8,895 samples in TCGA with Tissue-Agnostic and Tissue-Aware models, respectively, illustrating the mixture of two components: well-explained genes (green curve) and poorly-explained genes (red curve). The Tissue-Aware model demonstrates enhanced performance, with an average R^2 value of 0.70 compared to 0.45 for the Tissue-Agnostic model. **(C)** R^2 distribution of the linear models fitted using 1,000 landmark genes[7], displaying an average R^2 value of 0.77. The distribution exhibits a more pronounced bimodality, indicating that compared to the Tissue-Aware model, some genes are better explained, while the poorly-explained tail is also thicker. The Tissue-Aware model achieves comparable predictive power using only 56 predictors according to the R^2 measure and attains a correlation score of 0.9998 compared to 0.9975 for the L1000 genes (not shown), highlighting the effectiveness of the selected 56 regulatory molecules in predicting the transcriptome.

3.2 High Accuracy Tissue Classification with 56 Regulatory Molecules

Our analysis uses 56 regulatory molecules to perform multi-class tissue classification using a vanilla SVM classifier. Figure 2C illustrates the average confusion matrix obtained from 10-fold cross-validation, where the average accuracy of the classifier is 92.8%. However, the accuracy over tissue pairs is bimodal, indicating that the model is able to distinguish most pairs with 100% accuracy but struggles to classify certain pairs. Notably, challenging tissue pairs include COAD (Colon Adenocarcinoma) and READ (Rectal Adenocarcinoma), LUSC (Lung Squamous Cell Carcinoma) and LUAD (Lung Adenocarcinoma), and ESCA (Esophageal Carcinoma) and STAD (Stomach Adenocarcinoma). Additionally, higher-order classification confusion is observed for KIRP (Kidney Renal Papillary Cell Carcinoma), KIRC (Kidney Renal Clear Cell Carcinoma), and KICH (Kidney Chromophobe), which are all subtypes of renal cell carcinoma as well as squamous cancers including LUSC, CESC (Cervical Squamous Cell Carcinoma), HNSC (Head and Neck Squamous Cell Carcinomas), and BLCA (Bladder Carcinoma, with a basal-squamous subtype). The t-SNE plot of the clusters is shown in Figure 2B, with misclassified samples highlighted in yellow. The plot is zoomed in to emphasize that errors are more frequent in tissues that are highly similar. When using the TPM normalized data, the classification accuracy slightly decreased to 90.24%, though the misclassification patterns closely mirrored our observations with FPKM-UQ. These results are detailed in Supplement Figure S1C.

3.3 Highly Accurate Prediction of Gene Expression using Selected 56 Features

Our analysis employed the selected 56 features to predict the expression of approximately 20,000 genes using Tissue-Agnostic and Tissue-Aware models. Figure 3A and B illustrate the R^2 distribution histogram for both models, revealing that both histograms are mixtures of two components. The tail analysis procedure used Beta mixture distributions fitted to the

histograms, with the green curve representing the component covering the right tail (i.e., well-explained genes) and the red curve representing the left tail (i.e., poorly explained genes). Notably, the distribution of poorly-explained genes is thicker and skewed toward zero in the Tissue-Agnostic model in comparison to the Tissue-Aware model. Additionally, the variance explained by the model is significantly improved in the Tissue-Aware model. The average R^2 values for the Tissue-Agnostic and Tissue-Aware models are 0.45 and 0.70, respectively. When conducting similar experiments with TPM normalized data, we observed qualitatively similar histograms, presented in Supplement Figures S2A and B. Both histograms for the Tissue-Agnostic and Tissue-Aware models were identified as mixtures of two Beta distributions, with average R^2 values of 0.46 and 0.68, respectively.

3.4 Identifying Pathways Regulated by TFs, miRs, and Alternative Mechanisms

We obtained four sets of genes by using the 5th and 99th percentile cutoffs for poorly and well-explained genes in the tails of the R^2 distributions of the Tissue-Agnostic and Tissue-Aware models. In the Tissue-Agnostic model, we identified 256 poorly explained genes, while in the Tissue-Aware model, there were 106 poorly explained genes. Following gene set pathway enrichment, we found a high degree of overlap between enriched pathways that were considered significant. For both models, the number of significant pathways was limited to ten, with nine of them shared between the two models, as shown in the left column of Table 1. In the Tissue-Agnostic model, we identified 280 well-explained genes, while in the Tissue-Aware model, we identified 228 well-explained genes. After performing pathway enrichment analysis, we found 123 significantly enriched pathways in the Tissue-Agnostic model and 90 significant pathways in the Tissue-Aware model. Of these pathways, only 12 were significant in both models and are listed in the right column of Table 1.

Shared Pathways Enriched for	
Poorly-Explained Genes	**Well-Explained Genes**
Antimicrobial Peptides	Cell Adhesion Molecules (CAMs)
Beta Defensins	Hematopoietic Cell Lineage
Defensins	Tight Junction
Developmental Biology	Catecholamine Biosynthesis, Tyrosine =>Dopamine =>Noradrenaline => Adrenaline
GPCR Downstream Signaling	Hemostasis
Keratinization	MET Receptor Activation
Olfactory Signaling Pathway	Network map of SARS-CoV-2 Signaling J—pathway
Olfactory Transduction	Oligodendrocyte Specification and Differentiation, Leading to Myelin components for CNS Signaling by MSTL
Signaling by GPCR	Signaling by MSTL
	T Cytotoxic Cell Surface Molecules
	THelper Cell Surface Molecules
	Tight Junction Interactions

Table 1: **Pathway Enrichment Analysis for Poorly and Well-Explained Genes.** The table presents the results of gene set pathway enrichment analysis for genes poorly explained (left column) and well-explained (right column) common between Tissue-Agnostic and Tissue-Aware models. Out of 10 pathways enriched for genes poorly explained by both models, nine are shared, demonstrating that these pathways are not heavily regulated by TFs and miRNAs regardless of tissue type. Conversely, 123 and 90 pathways were enriched for well-explained genes in the Tissue-Agnostic and Aware models, respectively, among which only the listed 12 pathways (right column) were shared between the models. This indicates that tissue-specific regulatory mechanisms play a crucial role in governing the expression of well-explained genes. Poorly explained genes are potentially regulated by a broader range of factors. These might include chromatin accessibility, epigenetic modifications, RNA-binding proteins that alter the stability of transcripts, as well as other post-transcriptional events. On the other hand, well-explained genes are likely heavily regulated by TFs and miRNAs, participating in various cellular functions.

3.5 Comparison of Predictive Power of Selected Regulatory Molecules and L1000 Genes for Transcriptome

In this experiment, we compare the predictive power of our selected 56 regulatory molecules to that of the L1000 genes. Figure 3C shows the R^2 distribution of the linear model fitted using L1000 genes, with an average R^2 value of 0.77 over the entire transcriptome. In contrast, our Tissue-Aware model, using only 56 predictors, has an average R^2 value of 0.70, indicating comparable predictive power. The R^2 distribution of the L1000 genes also exhibits a similar bimodal shape as observed in the Tissue-Aware model. Furthermore, while the L1000 genes reach a correlation score of 0.9975, our Tissue-Aware model achieves a score of 0.9998. These results demonstrate that our selected 56 regulatory molecules are almost as good predictors of the transcriptome as the widely-used L1000 genes and, in some cases, even outperform them.

3.6 Distributional Shift and Mismatch of microRNAs Compromise Tissue-Aware Model's Transportability

To ensure the robustness and reliability of our Tissue-Aware model in predicting gene expression across different datasets, we investigated its generalization error [37] as a function of the number of Pseudo-Tissues. This error metric provides insights into the model's ability to capture underlying biological patterns in new, unseen datasets without merely fitting to known data. Figures 4A and B depict the cross-validation curves to tune the number of Pseudo-Tissues (clusters) for Root Mean Square Error (**RMSE**) and correlation score, respectively, using the FPKM-UQ normalized TCGA dataset. The optimal number of Pseudo-Tissues is found to be 15, based on both criteria. Additionally, we obtained analogous curves with the TPM normalized data, as shown in Supplement Figures S3A and B, which concur that 15 is the optimal number of Pseudo-Tissues.

Nevertheless, the transportability assessment (out-of-distribution generalization [38]) of the model on the CCLE dataset yielded a correlation score of 0.22 and RMSE of 3.83, which were worse than anticipated. We hypothesized that discrepancies in selected miRNAs and their distributional shifts [39] between TCGA and CCLE contribute to this diminished performance. To investigate this hypothesis, we assessed the distance of the distributions of matched features across TCGA and CCLE using Wasserstein distance [40], Figure 5. Specifically, the distance was calculated for each matched pair of features (either TF or miRNA), and a density plot was constructed based on these distances. For instance, the density for the 28 miRNAs stems from 28 paired TCGA and CCLE miRNA samples. This was then contrasted against a null distribution created from non-matching feature pairs. The findings, as illustrated in Figure 5A and B, revealed that while matched TFs in the datasets were more aligned than random pairs, the distances for matched miRNAs were comparable to those of random miRNA pairs. Such observations underscore the distinct distributional nature of the matched miRNAs, reinforcing the notion that a model trained on TCGA might encounter challenges when tested on CCLE.

The distance distribution offers a broad perspective on the discrepancies between the features of the two datasets. To further delve into individual variability of the features, we illustrated the distributions for three paired miRNAs in Figure 5C. Each of these pairs exemplifies a distinct miRNA matching challenge, as detailed in the Materials and Methods section. The unavailability of specific mature directional miRNA data in CCLE introduces potential mismatches, which can be considered as a technological source of covariate shift. This is shown by the histograms for hsa-miR-21a-5p (TCGA) versus hsa-miR-21a (CCLE) showcased in the top row of Figure 5C. Additionally, among the directionally matched miRNAs, certain ones displayed pronounced distributional shifts, indicative of a biological source of covariate shift. This is illustrated by hsa-miR-590-5p in the second row of Figure 5C. Moreover, in the absence of five crucial miRNAs in CCLE, they were paired with the most closely related alternatives, encapsulating both technological and biological sources of shift. This is represented in the third row of Figure 5C by hsa-miR-664a-3p (TCGA) paired with hsa-miR-29c (CCLE). Conversely, the TF distri-

butions exhibited remarkable consistency across the two datasets.

3.7 Robust Transportability of Tissue-Aware Model Using Transcription Factors Alone

Given the challenges we identified with miRNA consistency in our initial transportability experiments, we shifted our focus solely toward TFs. This decision was supported by the more stable expression distributions of TFs observed across datasets, as evident in Figure 5A. Utilizing 19 selected TFs from the TPM normalized TCGA dataset, we trained a Tissue-Aware model and subsequently evaluated its ability to predict the entire transcriptome on the CCLE, GTEx, and TARGET datasets. Remarkably, the model showcased superior performance on the CCLE when relying solely on TFs. The results were an average RMSE of 2.38 and a correlation score of 0.49. This was a notable improvement from the results obtained when both TFs and miRNAs were used as features.

Given the absence of separate miRNA measurements in the GTEx and TARGET datasets, we limited our model to TFs as predictors. The results were promising: for GTEx, we attained an average RMSE of 1.89 and a correlation score of 0.68. For TARGET, the corresponding values were an average RMSE of 2.34 and a correlation score of 0.53. These outcomes across the three distinct datasets underscore the transportability strength of our Tissue-Aware model. When relying exclusively on transcription factors, the model capably estimated the expression of approximately half of the genes across all target datasets.

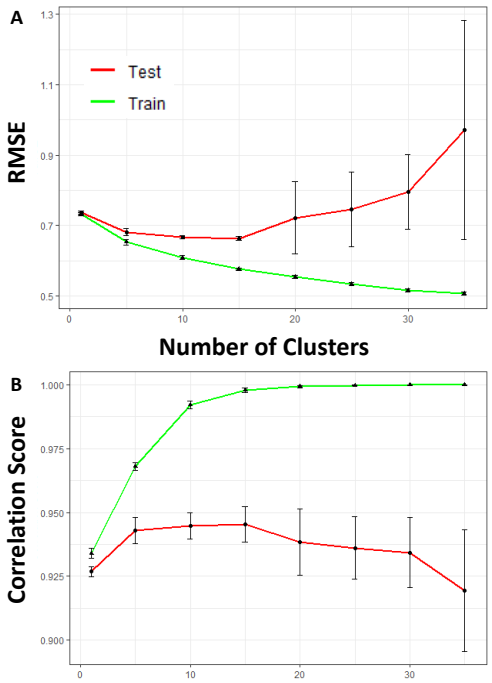

Figure 4: **Optimizing Pseudo-Tissues for Tissue-Aware Model Generalization. (A)** and **(B)** Cross-validation curves for tuning the number of Pseudo-Tissues (sample clusters) based on RMSE and correlation score, respectively, using the TCGA dataset. The optimal number of Pseudo-Tissues is determined to be 15 for both criteria. However, cross-validation assumes that the train and test data share the same distributions, which is not the case when distribution shifts are present, such as when selected miRNA expressions are heavily shifted in CCLE, thus hindering the transportability of the predictive model.

4 DISCUSSION

4.1 The Role of Causal Predictors in Robust Transcriptome Representation

Understanding gene regulatory networks, capturing cell states through gene expression profiles, and obtaining low-dimensional representations of cell states

are three essential and interconnected biological questions that, collectively, offer a cost-effective and robust method to evaluate the impact of various perturbations, encompassing mutations and therapeutic interventions. .

In this study, we demonstrated that by using our prior knowledge of the critical roles of transcription factors and microRNAs in gene regulation, we can establish a low-dimensional representation of cell states and infer the entire transcriptome from a limited number of regulatory molecules. The value of a reduced cell state representation lies in its ability to capture the gene expression distribution not only under normal conditions but also under various perturbations, such as drugs, mutations, or gene knockouts. These perturbations fundamentally change the joint distributions of gene expressions, generating interventional distributions [41] over expressions that can differ significantly from the normal state. To obtain the minimum number of genes needed to explain most transcriptome variations, it is necessary to understand or infer the gene regulatory network and select master regulators to fully predict the outcomes of perturbations. We hypothesized that using TFs and miRNAs as a shortcut can work in place of full gene regulation inference because of their significant role in the regulation of various biological processes. Consequently, under any perturbation or environment, their values serve as robust predictors of the entire transcriptome. Our hypothesis is supported by two key findings. Firstly, our feature selection pipeline identified regulatory molecule clusters that have biological significance, as reported in [12, 13]. Secondly, the selected features demonstrate strong predictive capability in supervised learning tasks across different environments, such as different tissues, and under different perturbations, such as DNA aberrations in cancer cells. Together, these results provide further evidence in support of our hypothesis.

4.2 Challenges in Cancer Type Classification

Our findings demonstrate that these selected features can effectively distinguish between 31 cancer types, as illustrated by the t-SNE plot (Figure 2A). Moreover, the classification accuracy obtained by our vanilla SVM classifier is promising at 92.8%, although the model faces some challenges with specific tissue pairs and subtypes. Biologically, the misclassification of these tissues can be rationalized. COAD and READ are both very similar subtypes of colorectal cancer [42], while LUSC and LUAD are subtypes of lung cancer, and ESCA and STAD are both types of gastrointestinal cancers. A similar situation arises for KIRP, KIRC, and KICH, where all are subtypes of renal cell carcinoma that originate from different cell types in close proximity to the kidney.

4.3 Enhanced Gene Expression Prediction Using Selected Features and Tissue-Aware Models

The prediction of gene expression using the selected 56 features was highly accurate, with an average R^2 value of 0.45 for the Tissue-Agnostic model and 0.70 for the Tissue-Aware model. The Tissue-Aware model showed significant improvement over the Tissue-Agnostic model, suggesting that incorporating tissue-specific information into our multi-task learning framework can enhance gene expression prediction. There are two potential biological explanations for the effectiveness of the Tissue-Aware model. First, several other types of regulatory molecules, such as epigenetic modifiers [43] and chromatin remodeling complexes [44], play essential roles in gene regulation besides TFs and miRs. Some of these, like DNA methylation and histone modification, are crucial in determining cell identity and, ultimately, tissue type [45, 46]. Therefore, using tissue type can serve as a surrogate for these omitted regulators. Secondly, from a causal perspective, the interaction between regulatory molecules can cause the causal effect of TFs and miRNAs to differ in each environment. Consequently, our multi-task learning framework, which fits a model per environment with shared parameters across environments, is a more biologically plausible model.

We hypothesized that the causal relevance of TFs and miRs, along with our flexible multi-task learning framework, would enable a succinct representation of the transcriptome. We confirmed this hypothesis by demonstrating that our selected 56 features

are as predictive as the widely-used L1000 genes. The predictive power of our 56 regulatory molecules was nearly as good as the L1000 genes, according to both R^2 and the correlation score measures. Our results highlight the effectiveness of our feature selection method, which is guided by biological knowledge about the causal role of regulatory molecules in transcription.

4.4 Pipeline Robustness to Diverse RNA-Seq Normalization Methods

In our analysis, we employed two well-established and commonly used normalization methods: FPKM and TPM. All steps of our study, with the exception of the transportability experiment, were conducted utilizing data normalized by both techniques. The primary distinction observed was in feature selection. Specifically, Thresher, our feature selection method, identified 19 TF clusters with TPM normalization, in contrast to the 28 identified using FPKM. However, the number of determined miRNA clusters remained consistent at 28 for both methods. While there was a 16% reduction in the total number of selected features ((28-19)/56) when employing TPM instead of FPKM, this reduction did not result in a significant performance drop in tissue type classification or transcriptome prediction. These observations underscore the robustness of our pipeline, irrespective of the chosen normalization method.

4.5 Factors Limiting the Prediction of Poorly-Explained Genes

In our analysis, a notable overlap (9/11=81%) emerges between the pathways enriched with genes that both the Tissue-Agnostic and Tissue-Aware models struggle to predict. This convergence prompts us to consider three potential reasons that might account for these similarities:

4.5.1 Alternative Regulation in Poorly Explained Genes.

Genes poorly explained by both Tissue-Agnostic and Aware models may not be heavily regulated by TFs and miRs. Gene enrichment analysis (left column of Table 1) indicates that these genes might be regulated by other mechanisms, such as signaling pathways [47], post-translational modifications [48], protein-protein interactions [49], and epigenetic regulation [45]. These alternative mechanisms may allow rapid and dynamic control of cellular responses, immune defense, and maintenance of cellular integrity. Examples of such pathways include antimicrobial peptides, keratinization, olfactory signaling, and GPCR downstream signaling.

4.5.2 Omission of Tissue-Specific microRNAs.

Another plausible explanation for our model's difficulty in accounting for certain poorly explained genes stems from our exclusion of microRNAs that have zero expression in 75% of the samples. Given that the expression and function of some miRNAs can be tissue-specific, it is possible that those we omitted might serve pivotal roles in regulating genes within those tissues. Thus, omitting them could hinder the model's capacity to adequately explain some genes.

4.5.3 Low Transcriptomic Presence of Poorly Predicted Genes.

Further analysis reveals that the average expression of genes that our model struggles to predict consistently ranks in the bottom 6% when compared to the average expression of all genes. This is further affirmed by the enrichment results; for instance, defensins, known for their specialized anti-microbial roles, are overrepresented among the poorly predicted genes. Due to their context-specific, low-level expression, our model may lack the sensitivity required for accurate prediction of these genes.

4.6 TF and miRNA Regulatory Roles in Well Explained Genes.

Genes well explained by both models are likely heavily regulated by TFs and miRNAs. Gene enrichment analysis (right column of Table 1) supports this hypothesis, showing that these genes are involved in various cellular functions, including cell adhesion, differentiation, neurotransmission, immune response, and barrier integrity. Tight regulation by

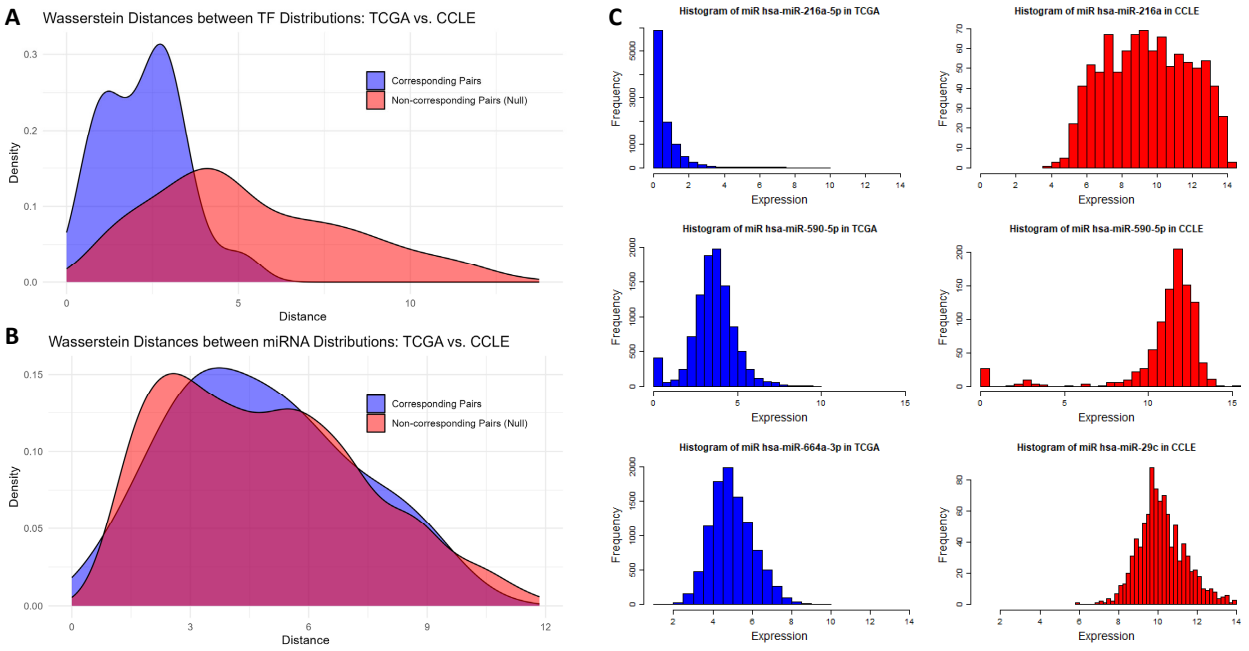

Figure 5: **Shift in Distributions of Regulatory Molecules across TCGA and CCLE Affecting Model Transportability.** The Tissue-Aware model's transportability challenges, when incorporating both TFs and miRNAs, contrast its efficacy with TFs alone, pointing toward differential miRNA distributions between TCGA and CCLE. The visualizations below substantiate this observation. **(A) Comparing TF Distributions.** The blue density visualizes the Wasserstein distance between corresponding (matched) TFs from TCGA and CCLE. It is shifted leftwards relative to the empirical null (red density). The null model is derived using distances from mismatched TF pairs across datasets. The data suggest TF distributions are predominantly congruent between TCGA and CCLE. **(B) Comparing miRNA Distributions.** For miRNAs, the landscape differs. The blue density, signifying the Wasserstein distance for matched miRNAs, aligns with the null (red density), indicating distinct miRNA distributions between the two datasets. **(C) Dissecting Cross Dataset microRNA Expression Variations.** Each row showcases a pair, with blue histograms representing TCGA and red denoting CCLE. The directionality of miRNAs in TCGA's may not be captured in CCLE, highlighted by hsa-miR-21a-5p (TCGA) versus hsa-miR-21a (CCLE) in row 1. Although certain miRNAs have their directionality recorded in both datasets, their distribution can still be different due to various biological or technological factors, illustrated by hsa-miR-590-5p in row 2. In five cases, we could not find the miRNAs of TCGA in CCLE. Our method then opted for the closest alternative, as seen with hsa-miR-664a-3p (TCGA) paired with hsa-miR-29c (CCLE) in row 3.

miRNAs and TFs allows cells to respond appropriately to stimuli [50] and maintain tissue homeostasis [51, 52], development [53], and function. Examples of such pathways include cell adhesion molecules (CAMs), hematopoietic cell lineage, tight junctions, catecholamine biosynthesis, and hemostasis.

A mere 5% overlap between the pathways enriched with genes accurately predicted by both the Tissue-Agnostic and Tissue-Aware models implies a distinctive capability of the Tissue-Aware model. It enhances the predictive accuracy for a multitude of genes, pushing them into the well-explained category (surpassing the 99th percentile threshold in R^2 distributions), thereby altering the set of pathways that are considered well-explained.

4.7 Impact of Covariate Shift on Transportability

The transportability of the Tissue-Aware model was tested on the CCLE dataset. Despite the promising performance on the TCGA dataset, covariate shift and mismatches, especially in microRNA expressions, affected the model's prediction accuracy on the CCLE dataset. Both the global picture of the shift in miRNAs as depicted in Figure 5B and individual examples illustrated in Figure 5C support the idea that the shift in miRNA distributions is significant and they are a culprit in the lack of transportability of our model. By focusing solely on TFs in our final experiment, the model's transportability performance significantly improved, confirming the detrimental effects of including mismatched or shifted miRNAs in the predictions. This observation emphasizes the necessity to ensure the robustness of features across diverse datasets to improve model transportability. The pronounced discrepancies in microRNA expressions between the datasets could be attributed to variations in experimental methodologies, preprocessing steps, or inherent biological differences.

4.8 Significance of Improved Transportability with the TF-only Tissue-Aware Model

The marked improvement in transportability, achieved when solely employing TFs for predictions, underscores the importance of transcription factors. As primary regulators of gene expression, they appear to offer a more stable and consistent feature set across diverse datasets. Yet, it is worth noting that the transportability performance does not match the training and testing outcomes from the TCGA datasets, suggesting opportunities for further refinement. Numerous potential research directions present themselves, which could enhance transportability. These include implementing domain adaptation techniques [39] to address covariate shifts in TFs or tapping into methods from multi-output prediction literature [54], where harnessing the inherent correlations between outcomes, here, the predicted gene expressions, there is a tangible opportunity to elevate overall prediction accuracy.

5 CONCLUSION

In this study, we demonstrated that using a small subset of transcription factors and microRNAs as selected features can effectively represent cell states and predict gene expression across 31 cancer types. Our results support the hypothesis that these regulatory molecules, due to their causal roles in gene regulation, can provide robust predictors of the transcriptome. The Tissue-Aware model, which incorporates tissue-specific information, outperformed the Tissue-Agnostic model, suggesting the importance of context for gene regulation. While our selected features were nearly as predictive as the L1000 genes, the model's transportability was affected by the covariate shift, particularly in microRNA expression. By using only TFs in the model, we improved our performance on the CCLE dataset, emphasizing the need for robust feature selection across different datasets. Future studies should explore additional regulatory molecules, such as epigenetic modifiers and chromatin remodeling complexes, as well as methods from domain adaptation and multi-output prediction

literature, to enhance the understanding of gene regulation and improve model transportability.

6 DATA AVAILABILITY

All the code and data used for analysis can be found on GitHub at -`https://github.com/aalabmaster/transcriptome_disentangled`.

7 ACKNOWLEDGEMENTS

Author contributions: Conceptualization, K.R.C.; methodology and investigation, A.A., K.R.C, Z.B.A.; writing, A.A; expert review and feedback on methods and findings, H.H.P.

8 ACKNOWLEDGEMENTS

NIH [R00HG011367 to A.A. and R21AI156292 to H.H.P].

8.0.1 Conflict of interest statement.

None declared.

References

[1] Mark Richards, Siew-Peng Tan, Jee-Hian Tan, Woon-Khiong Chan, and Ariff Bongso. The transcriptome profile of human embryonic stem cells as defined by sage. *Stem cells*, 22(1):51–64, 2004.

[2] Margarida Cardoso-Moreira, Jean Halbert, Delphine Valloton, Britta Velten, Chunyan Chen, Yi Shao, Angélica Liechti, Kelly Ascenção, Coralie Rummel, Svetlana Ovchinnikova, et al. Gene expression across mammalian organ development. *Nature*, 571(7766):505–509, 2019.

[3] Glenn A Maston, Sara K Evans, and Michael R Green. Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, 7:29–59, 2006.

[4] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.

[5] Marta Melé, Pedro G Ferreira, Ferran Reverter, David S DeLuca, Jean Monlong, Michael Sammeth, Taylor R Young, Jakob M Goldmann, Dmitri D Pervouchine, Timothy J Sullivan, et al. The human transcriptome across tissues and individuals. *Science*, 348(6235):660–665, 2015.

[6] John C Marioni, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9):1509–1517, 2008.

[7] Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452, 2017.

[8] Stephan Pabinger, Stefan Rödiger, Albert Kriegner, Klemens Vierlinger, and Andreas Weinhäusel. A survey of tools for the analysis of quantitative pcr (qpcr) data. *Biomolecular Detection and Quantification*, 1(1):23–33, 2014.

[9] Justin Lamb, Emily D Crawford, David Peck, Joshua W Modell, Irene C Blat, Matthew J Wrobel, Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N Ross, et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *science*, 313(5795):1929–1935, 2006.

[10] Wen Fury, Franak Batliwalla, Peter K Gregersen, and Wentian Li. Overlapping probabilities of top ranking gene lists, hypergeometric distribution, and stringency of gene selection criterion. In *2006 international conference of the IEEE engineering in medicine*

and biology society, pages 5531–5534. IEEE, 2006.

[11] Samuel A Lambert, Arttu Jolma, Laura F Campitelli, Pratyush K Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R Hughes, and Matthew T Weirauch. The human transcription factors. *Cell*, 172(4):650–665, 2018.

[12] Zachary B Abrams, Mark Zucker, Min Wang, Amir Asiaee Taheri, Lynne V Abruzzo, and Kevin R Coombes. Thirty biologically interpretable clusters of transcription factors distinguish cancer type. *BMC genomics*, 19:1–14, 2018.

[13] Amir Asiaee, Zachary B Abrams, Samantha Nakayiza, Deepa Sampath, and Kevin R Coombes. Explaining gene expression using twenty-one micrornas. *Journal of Computational Biology*, 27(7):1157–1170, 2020.

[14] Cancer Genome Atlas Research Network, John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas Pan-Cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

[15] Bo Li and Colin N Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12:1–16, 2011.

[16] Debra L Fulton, Saravanan Sundararajan, Gwenael Badis, Timothy R Hughes, Wyeth W Wasserman, Jared C Roach, and Rob Sladek. Tfcat: the curated catalog of mouse and human transcription factors. *Genome biology*, 10:1–14, 2009.

[17] GTEx Consortium, Kristin G Ardlie, David S Deluca, Ayellet V Segrè, Timothy J Sullivan, Taylor R Young, Ellen T Gelfand, Casandra A Trowbridge, Julian B Maller, Taru Tukiainen, et al. The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[18] Xiaotu Ma, YU Liu, Yanling Liu, Ludmil B Alexandrov, Michael N Edmonson, Charles Gawad, Xin Zhou, Yongjin Li, Michael C Rusch, John Easton, et al. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. *Nature*, 555(7696):371–376, 2018.

[19] Mahmoud Ghandi, Franklin W Huang, Judit Jané-Valbuena, Gregory V Kryukov, Christopher C Lo, E Robert McDonald III, Jordi Barretina, Ellen T Gelfand, Craig M Bielski, Haoxin Li, et al. Next-generation characterization of the cancer cell line encyclopedia. *Nature*, 569(7757):503–508, 2019.

[20] Mary J Goldman, Brian Craft, Mim Hastie, Kristupas Repečka, Fran McDade, Akhil Kamath, Ayan Banerjee, Yunhai Luo, Dave Rogers, Angela N Brooks, et al. Visualizing and interpreting cancer genomics data via the xena platform. *Nature biotechnology*, 38(6):675–678, 2020.

[21] Aviad Tsherniak, Francisca Vazquez, Phil G Montgomery, Barbara A Weir, Gregory Kryukov, Glenn S Cowley, Stanley Gill, William F Harrington, Sasha Pantel, John M Krill-Burger, et al. Defining a cancer dependency map. *Cell*, 170(3):564–576, 2017.

[22] Michael Lenz, Franz-Josef Müller, Martin Zenke, and Andreas Schuppert. Principal components analysis and the reported low intrinsic dimensionality of gene expression microarray data. *Scientific reports*, 6(1):1–11, 2016.

[23] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[24] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[25] Min Wang, Zachary B Abrams, Steven M Kornblau, and Kevin R Coombes. Thresher: determining the number of clusters while removing outliers. *BMC bioinformatics*, 19:1–15, 2018.

[26] Philipp Auer and Daniel Gervini. Choosing principal components: a new graphical method based on bayesian model selection. *Communications in Statistics—Simulation and Computation®*, 37(5):962–977, 2008.

[27] Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, Suvrit Sra, and Greg Ridgeway. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(9), 2005.

[28] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[29] Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*, 2020.

[30] William S Noble. What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567, 2006.

[31] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2021.

[32] Amir Asiaee, Samet Oymak, Kevin R Coombes, and Arindam Banerjee. High dimensional data enrichment: Interpretable, fast, and data-efficient. *arXiv preprint arXiv:1806.04047*, 2018.

[33] Amir Asiaee, Samet Oymak, Kevin R Coombes, and Arindam Banerjee. Data enrichment: Multi-task learning in high dimension with theoretical guarantees. In *Adaptive and Multitask Learning Workshop at the ICML. IMLS, Long Beach, CA*, 2019.

[34] Jing Chen, Eric E Bardes, Bruce J Aronow, and Anil G Jegga. Toppgene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic acids research*, 37(suppl_2):W305–W311, 2009.

[35] Zachary B Abrams, Travis S Johnson, Kun Huang, Philip RO Payne, and Kevin Coombes. A protocol to evaluate rna sequencing normalization methods. *BMC bioinformatics*, 20(24):1–7, 2019.

[36] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[37] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[38] Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

[39] Pirmin Lemberger and Ivan Panico. A primer on domain adaptation. *arXiv preprint arXiv:2001.09994*, 2020.

[40] Victor M Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *Annual review of statistics and its application*, 6:405–431, 2019.

[41] Judea Pearl. *Causality*. Cambridge university press, 2009.

[42] Cancer Genome Atlas Network et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330, 2012.

[43] Yuan Cheng, Cai He, Manni Wang, Xuelei Ma, Fei Mo, Shengyong Yang, Junhong Han, and Xiawei Wei. Targeting epigenetic regulators for cancer therapy: mechanisms and advances in clinical trials. *Signal transduction and targeted therapy*, 4(1):62, 2019.

[44] Chun Zhang, Jie Lu, and Pingyu Zhang. The roles of chromatin remodeling proteins in cancer. *Current Protein and Peptide Science*, 17(5):446–454, 2016.

[45] Bradley E Bernstein, Alexander Meissner, and Eric S Lander. The mammalian epigenome. *Cell*, 128(4):669–681, 2007.

[46] Howard Cedar and Yehudit Bergman. Linking dna methylation and histone modification: patterns and paradigms. *Nature Reviews Genetics*, 10(5):295–304, 2009.

[47] Jignesh R Parikh, Bertram Klinger, Yu Xia, Jarrod A Marto, and Nils Blï¿ 1/2thgen. Discovering causal signaling pathways through gene-expression patterns. *Nucleic acids research*, 38(suppl_2):W109–W117, 2010.

[48] Boxuan Simen Zhao, Ian A Roundtree, and Chuan He. Post-transcriptional gene regulation by mrna modifications. *Nature reviews Molecular cell biology*, 18(1):31–42, 2017.

[49] Safir Khan et al. Review on gene regulation: Dna-protein and protein-protein interactions and their regulatory elements. *Journal of Chemistry and Nutritional Biochemistry*, 2(2):35–45, 2021.

[50] Masafumi Inui, Graziano Martello, and Stefano Piccolo. Microrna control of signal transduction. *Nature reviews Molecular cell biology*, 11(4):252–263, 2010.

[51] Astrid Eijkelenboom and Boudewijn MT Burgering. Foxos: signalling integrators for homeostasis maintenance. *Nature reviews Molecular cell biology*, 14(2):83–97, 2013.

[52] Karthikeyan Chandrasekaran, Dwi S Karolina, Sugunavathi Sepramaniam, Arunmozhiarasi Armugam, E Marelyn Wintour, John F Bertram, and Kandiah Jeyaseelan. Role of micrornas in kidney homeostasis and disease. *Kidney international*, 81(7):617–627, 2012.

[53] François Spitz and Eileen EM Furlong. Transcription factors: from enhancer binding to developmental control. *Nature reviews genetics*, 13(9):613–626, 2012.

[54] Hanen Borchani, Gherardo Varando, Concha Bielza, and Pedro Larranaga. A survey on multi-output regression. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(5):216–233, 2015.

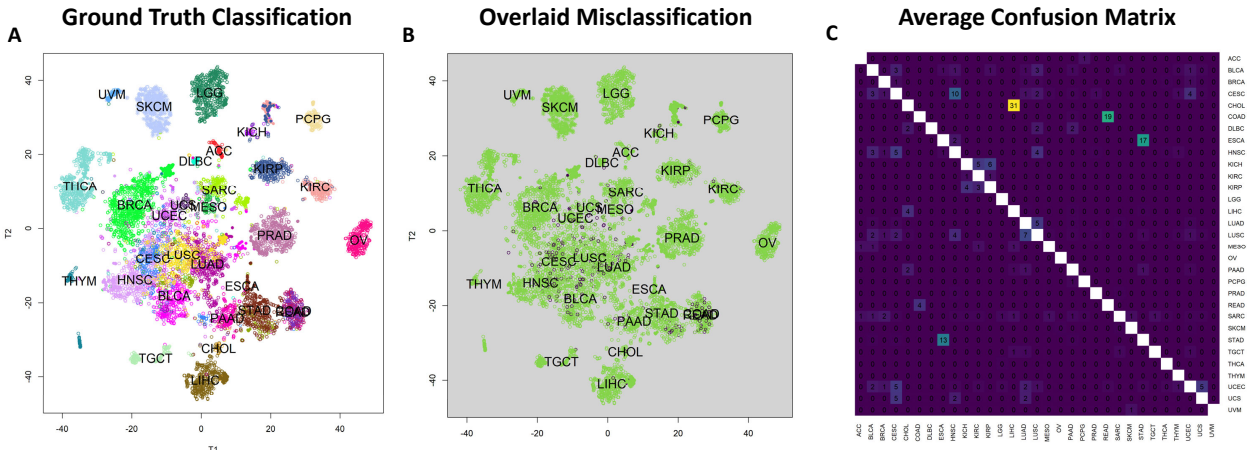

Figure S1: **Cancer Type Classification and Visualization Through Regulatory Molecules for TPM normalized TCGA Data. (A)** A 2D t-SNE plot of the 8895x47 input TF and miRNA expression matrix, consisting of selected 19 TFs and 28 miRNAs, with colors corresponding to known cancer types, showcasing the ability to differentiate most cancers using 47 regulatory molecules. **(B)** An enlarged t-SNE plot focusing on misclassified samples, marked in dark purple, indicates that classification errors are more common in highly similar cancer types. **(C)** An average confusion matrix generated from 10-fold cross-validation, exposing the error pattern of the SVM classifier with 90.24% accuracy. The numbers within each cell of the heatmap represent the average percentage of times, across 10-fold cross-validation, that samples from one class (row) were predicted to belong to another class (column). A value of 100% would indicate that all samples of a given class were consistently misclassified as another specific class across all cross-validation folds. Diagonal cells representing correct classifications have been left blank for clarity. Lighter matrix entries (excluding diagonals) represent larger errors, and the error pattern predominantly coincides with the t-SNE visualization of errors in panel B. Classifying similar cancer pairs like STAD and ESCA, COAD and READ, LUSC and LUAD, higher-order similar cancers such as KIRP, KIRC, and KICH, as well as squamous cancers (or cancers with a squamous cell subtype) including LUSC, CESC, HNSC, and BLCA, is challenging. All results are similar to FPKM normalized data with 56 features.

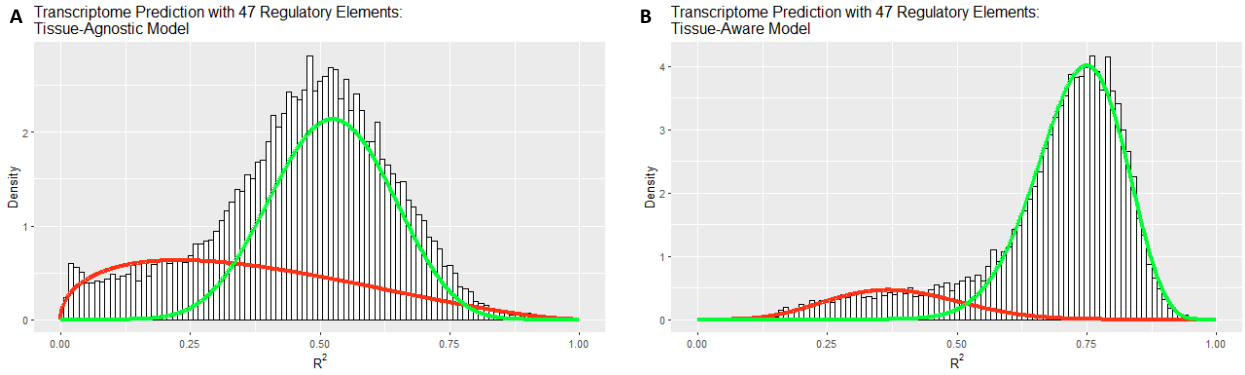

Figure S2: **Gene Expression Prediction of Selected Regulatory Molecules with TPM Normalized Data. (A)** and **(B)** R^2 histograms for predicting 20,289 gene expressions using 8,895 samples in TCGA with Tissue-Agnostic and Tissue-Aware models, respectively, illustrating the mixture of two components: well-explained genes (green curve) and poorly-explained genes (red curve). The Tissue-Aware model demonstrates enhanced performance, with an average R^2 value of 0.68 compared to 0.46 for the Tissue-Agnostic model.

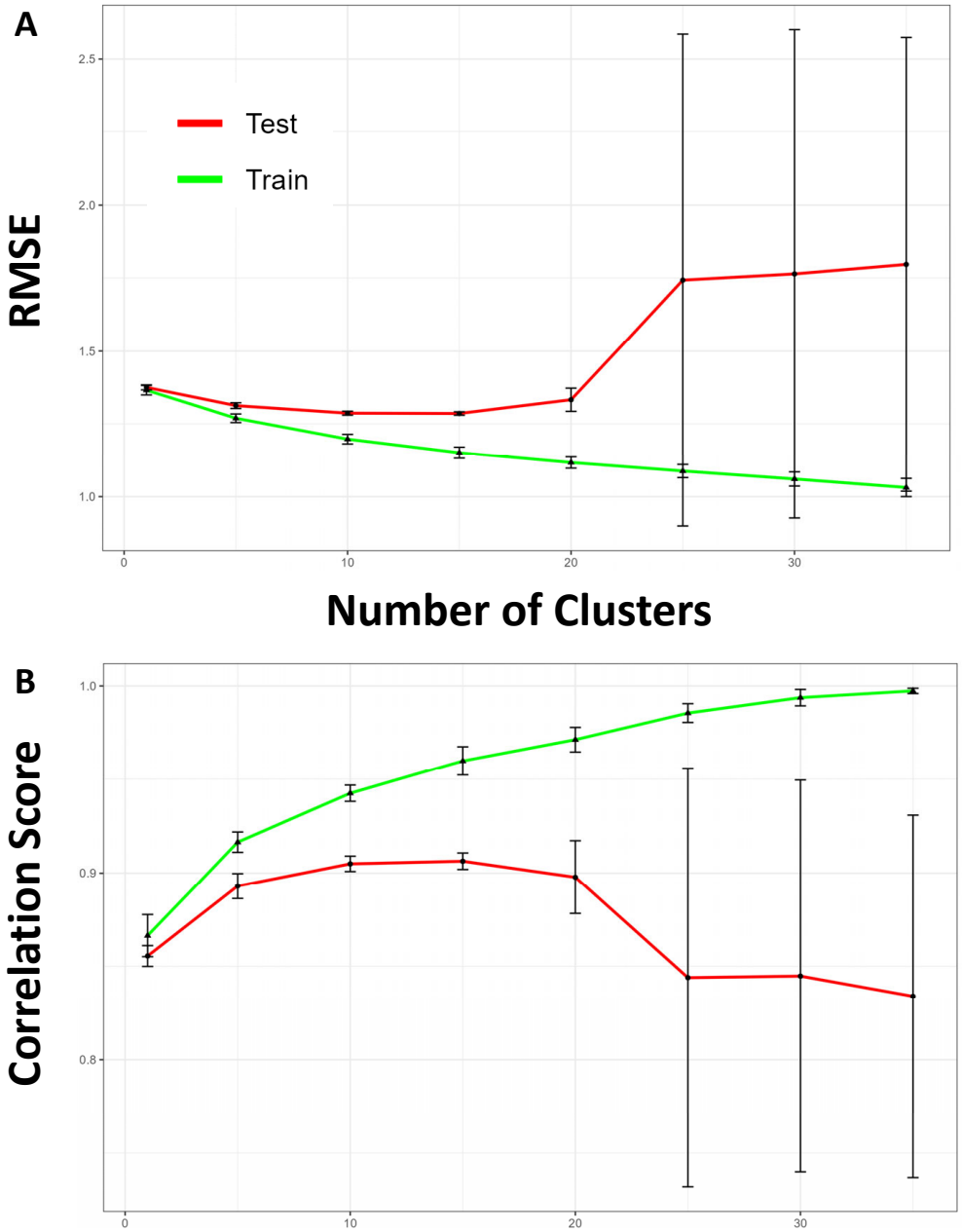

Figure S3: **Optimizing Pseudo-Tissues for Tissue-Aware Model Generalization using TCGA TPM Normalized Data. (A)** and **(B)** Cross-validation curves for tuning the number of Pseudo-Tissues (sample clusters) based on RMSE and correlation score, respectively, using the TCGA dataset. The optimal number of Pseudo-Tissues is determined to be 15 for both criteria. However, cross-validation assumes that the train and test data share the same distributions, which is not the case when distribution shifts are present, such as when selected miRNA expressions are heavily shifted in CCLE, thus hindering the transportability of the predictive model.