Department of Physiology and Pharmacology
Karolinska Institutet, Stockholm, Sweden

MODELING TRANSCRIPTION FACTOR REGULATION USING DEEP LEARNING ON BULK AND SINGLE-CELL RNA-SEQ DATA

Kejun Li

Karolinska Institutet

Stockholm 2025

Modeling Transcription Factor Regulation Using Deep Learning on Bulk and Single-Cell RNA-seq Data

Thesis for Master's Degree (MSc) in Translational Physiology and Pharmacology

By

Kejun Li

The thesis will be defended in public at Stockholm, 3nd June 2025.

Supervisor:
Cheng Zhang
KTH Royal Institute of Technology
Division of SYSTEMS BIOLOGY

Acknowledgement of the use of generative AI.
During the preparation of this work, the author(s) used ChatGPT (version GPT-4), OpenAI, https://openai.com/chatgpt to assist in refining technical language and summarizing complex model architecture.
After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the work.

Popular science summary of the thesis

Every cell in our body contains the same DNA, yet cells in the brain behave very differently from those in the liver or the skin. One reason for this is the action of transcription factors—proteins that control which genes are turned on or off in a particular cell. Understanding how transcription factors regulate gene activity is key to uncovering how our cells work, and what goes wrong in diseases.

However, figuring out these gene control networks is extremely complex. Traditional laboratory methods are time-consuming and expensive, and they don't work well across different cell types. To address this, researchers are increasingly turning to artificial intelligence (AI) and deep learning, a form of machine learning inspired by how the brain processes information.

In this study, we developed a deep learning model that can predict gene activity based on the presence of transcription factors in a cell. What makes our model special is that it combines AI with prior biological knowledge—specifically, information about where transcription factors are likely to bind in the genome.

We trained our model using large-scale genetic data from human liver and kidney tissues. By comparing different AI techniques, we found that a type of model called a recurrent neural network (RNN) gave the best results. Our model could predict gene activity patterns in both bulk tissue samples and, to a lesser extent, in individual cells. It also allowed us to explore which transcription factors play key roles in regulating gene networks.

While promising, the model also revealed challenges—for example, predicting gene activity in highly varied single-cell data was difficult, and the model sometimes struggled with genes that are usually inactive. Future improvements might include better handling of rare cell types or using protein activity data to get a more accurate picture.

Abstract

Understanding transcription factor (TF) regulation is critical for elucidating the gene expression programs that drive cellular function and disease. Traditional experimental and statistical approaches face challenges such as data sparsity, condition-specificity, and limited interpretability. In this study, we present a deep learning framework that predicts gene expression patterns by integrating TF motif-based prior knowledge with transcriptomic data from both bulk and single-cell RNA sequencing. Our model constructs a regulatory network using TF-DNA motif mappings and employs a recurrent neural network (RNN) architecture to learn edge-specific regulatory weights from expression data. We demonstrate that the RNN model outperforms traditional machine learning and graph neural network approaches in both predictive accuracy and generalization. Evaluations across liver and kidney datasets show robust performance, with particularly strong results in bulk tissue data. Furthermore, the model enables mechanistic interpretation by identifying key regulatory factors and potential novel interactions, many of which are supported by existing biological databases and literature. This work highlights the power of biologically informed deep learning in modeling gene regulation and offers a scalable framework for integrative transcriptomic analysis across diverse biological contexts.

Contents

List of abbreviations

RNN	Recurrent Neural Network
TF	Transcription Factor
CNN	convolutional neural networks

1. Introduction

In eukaryotic cells, transcription factors (TFs) regulate target gene expression by recognizing specific DNA sequences, thereby establishing complex gene regulatory networks. These networks play pivotal roles in cellular differentiation, tissue homeostasis maintenance, and disease progression. Accurately deciphering the regulatory relationships between TFs and their target genes is crucial for understanding cellular functions and regulatory mechanisms. Traditional network construction methodologies predominantly rely on ChIP-seq and gene perturbation experiments, or statistical inference based on co-expression patterns. However, experimental data exhibit limitations in coverage and condition specificity, while expression data-based modeling faces challenges of high-dimensional noise, data sparsity, and regulatory hierarchy ambiguity.

In recent years, deep learning has been extensively applied to various bioinformatics tasks due to its robust capability in modeling high-dimensional nonlinear relationships, including TF-DNA binding site prediction, chromatin state analysis, and enhancer activity identification. Particularly in TF binding prediction, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been employed to extract features from DNA sequences, demonstrating exceptional performance. In enhancer recognition, models such as DeepEnhancerPPO, which integrates ResNet with Transformer architecture and incorporates a reinforcement learning strategy, have improved classification performance while enhancing model interpretability. However, most current deep learning models still primarily depend on sequence features or directly fit expression data through fully connected structures, with limited systematic integration of structural prior information from transcriptional regulatory networks, thereby constraining interpretability and generalization capabilities. Several studies have indicated inconsistent performance of traditional models across different cell types, with difficulties in elucidating underlying regulatory mechanisms.

In this study, we develop a deep learning framework that predicts gene expression patterns based on transcription factor activity, incorporating prior knowledge derived from TF motif mapping to regulatory DNA elements. Specifically, we construct a transcriptional regulatory mega network using TF motif information and apply an RNN model that learns edge-specific regulatory weights from gene expression data. This approach enables us to predict gene expression and extract mechanistic hypotheses about transcriptional regulation.

By evaluating our model across both bulk and single-cell RNA-seq datasets from liver and kidney tissues, we assess its performance, robustness, and generalizability. Our findings offer insights into regulatory architecture and highlight the strengths and limitations of biologically informed deep learning models in genomic research.

2. Research aims

The overarching aim of this research is to develop a deep learning approach for predicting gene expression based on transcription factor regulation, using bulk or single-cell RNA-seq data. The specific objectives are:

1. Build a network of potential regulatory interactions between transcription factors and target genes. The network serves as prior knowledge to constrain the modeling process.

2. Develop deep-learning models that map TF expression vectors to target-gene expression and capture edge-specific regulatory weights indicating activation or repression.

3. Evaluate model performance via internal cross-validation experiments and independent external datasets.

4. Query the learned network weights to derive mechanistic hypotheses regarding the regulatory actions of transcription factors on gene expression levels in hepatic tissue.

5. Provide an open, extensible framework for modelling transcriptional regulation that can be adapted to other tissues, conditions.

3. Materials and methods

Knowledge Network Construction

From the footprintDB database, we downloaded a collection of human transcription factor motifs, ultimately acquiring over 4,000 distinct DNA binding motifs encompassing approximately 1,200 transcription factors. We obtained the human reference genome sequence (GRCh38 primary assembly) and its regulatory region annotation data from Ensembl (version 113). Based on these annotations, we utilized the Bedtools suite to extract sequences for each regulatory region from the genomic files, which contained sequences from regulatory regions, encompassing potential transcription factor binding sites proximal to genes. The regulatory region sequences were scanned using FIMO v5.5.7 software from the MEME Suite toolkit, with a p-value threshold of 1e-5 to identify statistically significant motif matches. Bedtools was employed to map each motif match to Ensembl regulatory features, retaining only overlaps within promoters or enhancers. To construct the transcriptional regulatory network, we established TF-gene regulatory edges by linking each transcription factor to genes whose promoters or enhancers contained significant matches (P<1e-5?) to the TF's binding motif. If a motif corresponding to a TF was found within a regulatory region associated with a gene, an edge was created between the TF and the gene, thereby forming a motif-based regulatory network.

Expression Data Sources and Preprocessing

Bulk RNA sequencing data from human liver tissue, obtained from the ARCHS4 project, were used for model training. For external validation, we employed single-cell RNA sequencing data from human kidney cell and human liver tissue bulk RNA sequencing data. For bulk RNA sequencing data, raw counts were converted to Transcripts Per Million (TPM), followed by logarithmic transformation of the TPM values. For single-cell RNA sequencing data, cells expressing fewer than 200 genes and genes expressed in less than 10% of cells were excluded. The raw counts were normalized to 10,000 per cell and logarithmically transformed. The MAGIC imputation algorithm was applied to restore gene-gene correlations. During the modeling process, our primary focus was on the transcription factors and their target genes encompassed within the knowledge network.

Comparative Performance Analysis of Different Frameworks

To evaluate model performance, 900 samples were randomly selected from the liver tissue bulk sequencing data. This study compared the predictive efficacy of three models: Support Vector Regression (SVR), Recurrent Neural Network (RNN), and Graph Attention Network (GAT). All models employed identical data partitioning, specifically dividing the complete dataset into 80% training set and 20% test set. Both the RNN and GAT models utilized Mean Squared Error (MSE) as the loss function, with the initial output activation function set to MML(Michaelis–Menten-like) , and were trained under identical conditions regarding the best iteration count and learning rate. The model demonstrating superior predictive results was selected for comprehensive training and final model establishment.

Model Structure

This research proposes a Recurrent Neural Network (RNN) framework that predicts the expression of all genes based on the expression of transcription factor downstream target genes. The overall architecture consists of three sequentially connected sub-modules: an input projection module, a transcriptional regulatory network module, and an output projection module. The input projection module receives expression vector of transcription factor downstream target genes as input and maps it to the full node space of the transcriptional regulatory network through linear transformation. This module applies element-wise weighting to the input using a trainable parameter vector, scaled by a constant factor.

The core of the transcriptional regulatory network module is a recurrent neural network constructed based on prior network topology, where each node represents a transcription factor or gene, and each edge represents the regulatory relationship between a transcription factor and its target genes. Node state updates depend on the states of adjacent nodes from the previous time step and their connection weights, with all connection weights represented as a sparse adjacency matrix. This module implements non-linear updates using a MML that exhibits biological molecular saturation dynamics, and introduces a "leakage" parameter to enhance numerical stability. The transcriptional regulatory network iterates up to 300 times on each training sample, with convergence considered achieved when changes between node states fall below a set threshold (e.g., 1e-5). The output projection module extracts the final states of a predefined set of gene nodes from the transcriptional regulatory network and applies a linear transformation to predict the expression levels of target genes.

Model Training

The model is trained by minimizing the Mean Squared Error (MSE) between predicted gene expression and actual expression, supplemented with multiple regularization terms that reinforce biological prior constraints and enhance numerical stability. The training process employs the Adam optimizer over 600 iterations, with the learning rate dynamically adjusted between 2e-3 and 1e-8 according to training progress. To enhance the model's ability to recognize regulatory directions, a sign consistency penalty is introduced during training. Specifically, for edges with known interaction directions, an L1 penalty (coefficient of 0.1) is applied when the learned connection weights contradict prior knowledge. Additionally, L2 regularization (coefficient of 1e-6) is applied to all weight and bias parameters to prevent overfitting and constrain parameter ranges. To prevent weights from converging to zero values, the model further incorporates an auxiliary regularization term in the form of $1/(w^2 + 0.5)$ to increase gradient sensitivity. Regarding the dynamic distribution of node states, the model imposes constraints on the state distributions of transcription factors and gene nodes under different sample conditions. Specifically, node states sorted by sample are compared item by item with an ideal uniform distribution, and their mean squared error is calculated; penalties are applied to state values outside a preset range, thereby constructing a regularization loss to encourage reasonable node state distributions, with an intensity set at 1e-4. To ensure the stability of the transcriptional regulatory network, the model introduces spectral radius constraints. The maximum eigenvalue of the network under each batch of samples is approximated using random perturbation methods, and an exponential barrier function is employed to control it within the target range (<1), with a corresponding loss term coefficient of 1e-5. After each round of training, the weights of all non-interactive edges are reset to zero to maintain the consistency of the network structure.

Model Validation and Generalization

Five-fold cross-validation was employed for the internal validation process during model training, with random partitioning of training and validation sets in each fold without repetitive sampling. Upon completion of training, the resulting models were preserved and subsequently applied to each external validation dataset to generate predictions. The predictive performance was evaluated using global Pearson correlation coefficient (r), mean squared error (MSE), and coefficient of determination (R^2) as assessment metrics.

Transcription Factor-Target Gene Network Analysis

Following model training completion, we conducted a systematic structural analysis of the gene-transcription factor regulatory network derived from the model output. Initially, we extracted the weight matrix (transcription factors × target genes) from the trained model to construct a directed weighted regulatory network, which was subsequently transformed into an edge list for statistical analysis of weight distribution across all regulatory connections. High-weight edges were selected by establishing an upper quantile threshold (e.g., top 1%), defining regulatory relationships where transcription factors exerted maximal promotional effects on target gene expression. Subsequently, we employed NetworkX to construct the regulatory network graph and calculated various node centrality metrics, including Degree Centrality, Betweenness Centrality, and PageRank scores, to respectively identify transcription factors with extensive regulatory scope, critical intermediary nodes, and regulatory factors possessing global influence.

We applied the Louvain algorithm for community detection, treating each module's node collection as a candidate co-regulatory unit. Subsequently, gene sets within each module underwent enrichment analysis for GO terms (Biological Process) and pathway databases (e.g., KEGG, Reactome) to identify potential biological functions and signaling pathway involvement. For significantly enriched modules, we further analyzed potential signal activation processes or functional differentiation mechanisms.

We cross-referenced the predicted regulatory relationships and network structures against multiple biological databases. Specifically, high-intensity regulatory edges were compared with manually curated transcription factor-target gene regulatory relationships from the TRRUST database; concurrently, protein interaction information was obtained via the STRING API, and joint keywords for key regulatory pairs were queried in the PubMed database.

Hardware and software specifications

Our model was implemented and trained using Python (version 3.10.16) and built upon the PyTorch framework (version 2.4.1), with CUDA (version 11.8) acceleration. All simulations, evaluations, and cross-validation procedures were conducted on the Berzelius cluster at NSC, equipped with NVIDIA® DGX-A100 computational nodes.

4. Results

Constructed TF-Gene Network Characteristics

The final regulatory network constructed comprises 1,200 transcription factors and 16,292 target genes, interconnected by 202,550 directed edges, with each edge representing a potential regulatory relationship between a transcription factor and its target gene. Certain transcription factors were predicted to target hundreds of genes, typically due to the widespread presence of their binding motifs (such as SP1 and AP-1 family members) across multiple promoter regions; conversely, other transcription factors were associated with only a limited number of target genes, potentially reflecting their tissue-specific regulatory functions. It is important to note that this network likely encompasses both genuine functional regulatory relationships and a proportion of false positive predictions. To address this, our modeling framework retains all potential edges while allowing the model to assign weights to each edge based on actual expression data.

Model Performance on Training Data

Table 1. *Prediction performance from SVR, RNN, and GAT model framework on identical datasets*

Model	MSE	R^2	PCC
SVR	0.1421	-31.9967	0.8689
RNN	0.2002	0.6416	0.8025
GAT	0.3332	0.4036	0.6377

Model performance was evaluated using Mean Squared Error (MSE), coefficient of determination (R^2), and Pearson Correlation Coefficient (PCC), with detailed results presented in Table 1. Among the three models, SVR exhibited the lowest MSE (0.1421), indicating minimal average prediction error; concurrently, it achieved the highest PCC (0.8689), demonstrating a strong linear correlation between predicted and actual values. However, the SVR model yielded a significantly negative R^2 value (-31.9967), suggesting its explanatory capacity for data variance was substantially inferior to simple baseline models, potentially attributable to overfitting or insufficient generalization capability. In contrast, the RNN model demonstrated superior balance across all metrics. With an MSE of 0.2002, R^2 of 0.6416, and PCC reaching 0.8025, it demonstrated a robust capacity to explain target variable variance. Despite being trained under identical conditions as the RNN, the GAT model

underperformed: it exhibited the highest MSE (0.3332), lowest R^2 (0.4036), and weakest PCC (0.6377), indicating limited predictive capacity within this experimental framework.

Considering these comprehensive results, the RNN model demonstrated optimal equilibrium between prediction accuracy and generalization capability, consequently emerging as the preferred architectural choice for subsequent comprehensive training and ultimate model deployment.

External Validation on Independent Datasets

Following confirmation that the RNN exhibited optimal performance on the training data, we further evaluated its generalization capability on entirely independent datasets. Specifically, we tested the liver-data-trained model under four distinct contexts: bulk RNA sequencing of liver tissue, single-cell RNA sequencing of hepatocytes, bulk RNA sequencing of kidney tissue, and single-cell RNA sequencing of kidney cells. The evaluation results are presented in Table 2.

Table 2. *Validation performance of the RNN model on various datasets.*

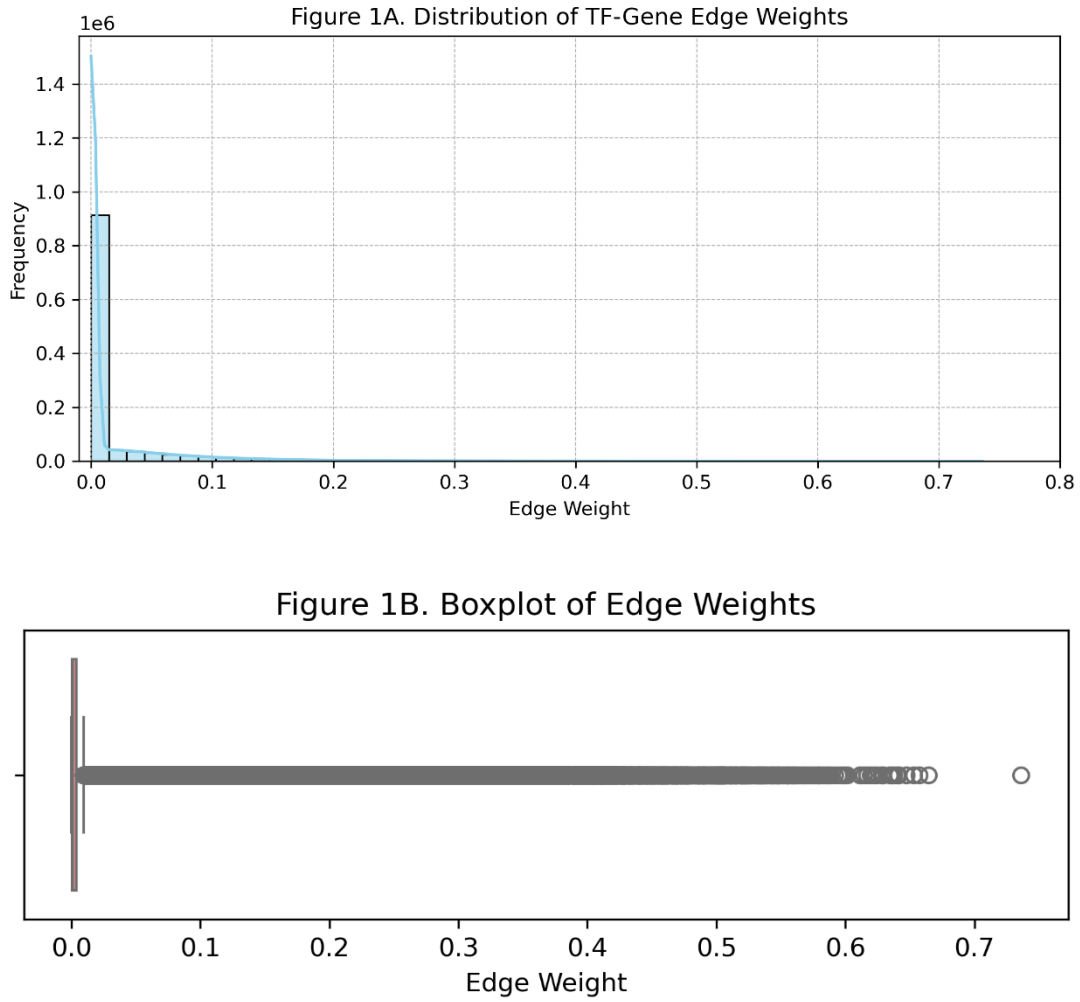
Dataset	MSE	R^2	PCC
Liver – cross validation (5-fold)	0.1162	0.5031	0.7553
Liver – bulk RNA-seq (262 samples)	0.0315	0.9063	0.9524
Liver – single-cell RNA-seq (4892 cells)	0.0611	0.3060	0.6823
Kidney – bulk RNA-seq (104 samples)	0.1359	0.6405	0.8211
Kidney – single-cell RNA-seq (21186 cells)	0.0427	–0.7122	0.4712

Network Construction and Edge Weight Distribution Analysis

Upon completion of model training, we extracted the output weight matrix (dimensions: number of transcription factors × number of target genes) to construct a directed weighted regulatory network. This network comprised millions of edges, representing the regulatory intensity of transcription factors on target genes. We transformed the edge weights into an edge list and analyzed their global distribution characteristics. Figure 1A demonstrates that most edge weights were extremely low, with only a small fraction exhibiting significantly high regulatory intensity, presenting a highly skewed distribution. This trend was further substantiated in the box plot (Figure 1B), where the vast majority of edge weights clustered in exceedingly low ranges, with only a few extreme values in the high-weight region. To

identify the most significant regulatory relationships, we designated the top 1% of edges as high-effective regulatory connections (Supplementary File 1) and utilized these for subsequent network structure and functional analyses. The high-effective regulatory connections encompass 818 distinct transcription factors, 6,511 different genes, and 11,540 transcription factor-gene interaction relationships.

Figure 1. *Distribution of regulatory edge weights in model output.*



Network Structure and Centrality Analysis

Employing the NetworkX toolkit, we structurally modeled the aforementioned regulatory network and calculated multiple node centrality metrics. Table 3 lists the top 10 transcription factors ranked by PageRank, which exhibit high global regulatory capacity within the network and potentially function as master regulators influencing multiple target gene modules.

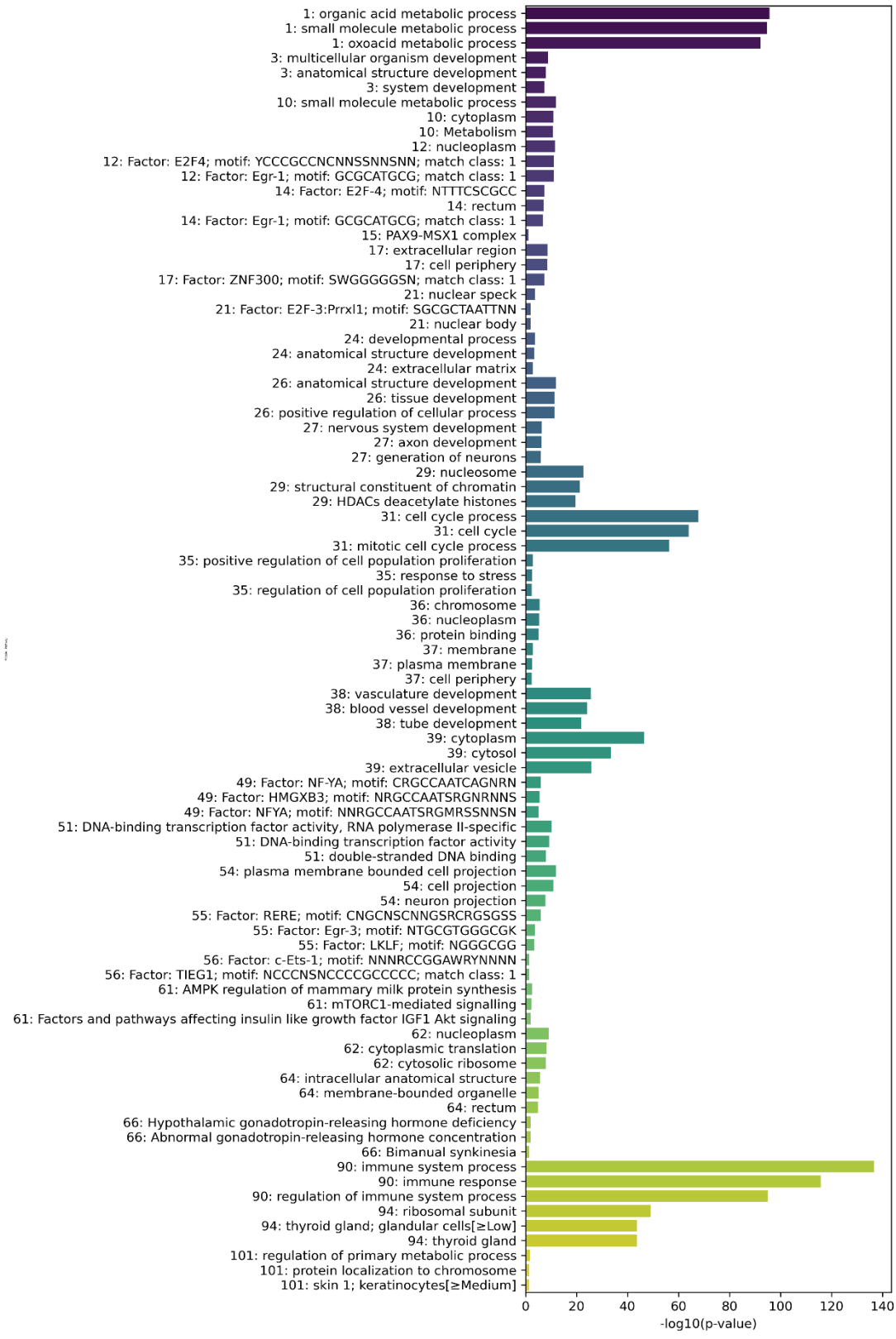
Table 3. *Genes with highest global influence in the network*

Node	Degree Centrality	PageRank
FTL	0.0052	0.000986
RPS27	0.0028	0.000664
SERPINA1	0.0025	0.000656
ATP5F1B	0.0020	0.000545
MYL12A	0.0011	0.000527
FGA	0.0014	0.000522
RPS12	0.0032	0.000449
GSC2	0.0003	0.000416
DBX1	0.0003	0.000416
ALB	0.0031	0.000408

Module Detection and Functional Enrichment Analysis

To identify potential co-regulatory substructures, we applied the Louvain algorithm for community detection on the regulatory network, partitioning it into multiple functional modules. We extracted gene sets from each module for enrichment analysis. Figure 2 illustrates significantly enriched biological processes within each module (ranked by -log10(p-value)). For instance, module 0 primarily exhibited enrichment in metabolism-related pathways (such as small molecule metabolism and oxidative acid metabolism), module 4 was associated with immune system regulation, while module 11 concentrated on cell cycle processes, suggesting distinct functional specialization across different modules.

Figure 2. *Most significant functional enrichment results for each module.*



Biological Validation and Regulatory Credibility Assessment

To validate the biological credibility of model-predicted regulatory relationships, we compared the selected high-weight TF-target gene pairs against known regulatory pairs documented in the TRRUST database, revealing numerous strong interactions consistent with database records, thus supporting the model's capacity to reproduce established biological

regulatory relationships. Additionally, we conducted literature mining via PubMed, searching for keyword co-occurrences among TF-target gene combinations with high PageRank scores. Multiple high-confidence regulatory edges were substantiated by explicit experimental evidence in the literature, further enhancing the reliability of the predictions. For detailed results, please refer to Supplementary File 2. Comparative analysis with the TRRUST database revealed that 334 of the identified regulatory relationships have been experimentally validated, demonstrating the model's capability to recognize classical regulatory interactions. Protein interaction data obtained through the STRING API identified 879 pairs as protein-protein interaction partners, supporting the existence of direct physical contacts between transcription factors and their target genes. Text mining via PubMed keyword co-occurrence searches indicated that 1,925 pairs have been previously reported in the literature, suggesting that the model's predictions are substantiated by existing research findings.

5. Discussion

In this study, we attempted to model gene expression driven by TFs using deep learning approaches, incorporating prior knowledge of TF-gene interactions derived from sequence motif analysis. Our results reveal both the potential and challenges of this methodology.

A significant finding is that RNN models substantially outperformed GAT in our experimental setting. Specifically, RNNs achieved high predictive accuracy in cross-validation on liver data ($R^2 > 0.8$), while GATs performed relatively poorly ($R^2 \sim 0.4$). Although GATs are theoretically well-suited for graph-structured problems and have demonstrated strong performance in other domains, this performance disparity may stem from multiple factors. First, RNNs are designed to share parameters across all genes, substantially reducing the number of free parameters the model needs to fit; in contrast, GATs require learning at least one attention weight for each edge or node pair. Given our limited training samples, RNNs may generalize more effectively, whereas GATs are more susceptible to insufficient training or overfitting. Second, gene regulatory processes may possess sequential or cumulative characteristics, where gene expression might reflect the additive influence of multiple TF inputs. RNNs can capture such cumulative relationships through sequential input processing, while GATs employ parallel attention weighting that may struggle to model nonlinear cumulative effects. Although GAT performance might improve through more refined hyperparameter tuning or larger training datasets, RNNs unquestionably demonstrated superior efficiency and effectiveness in this study. This result also emphasizes that simpler structures with fewer parameters may outperform more complex architectures in biological deep learning scenarios with limited data.

External validation further indicates that the training data context is crucial for model generalization. RNN models trained on liver data performed excellently on other liver bulk samples, suggesting they captured genuine regulatory signals rather than data noise. However, while the model reflected partial trends in liver single-cell data, it struggled to encompass cellular heterogeneity, revealing fundamental differences between bulk sample averages and single-cell variations. Model performance on kidney bulk data was adequate, indicating that some regulatory mechanisms may be shared across tissues, or that the model learned cross-tissue applicable basic regulatory programs. For example, fundamental processes like ribosome biogenesis and energy metabolism are often regulated by identical

TFs across different cell types; the model may have learned these patterns from liver data and successfully transferred them to kidney, thus achieving good predictive performance for many housekeeping genes (as evidenced by higher Pearson correlation coefficients). However, performance on kidney single-cell data was almost unsuccessful, demonstrating that simultaneously extrapolating to new tissues and new data modalities remains highly challenging. This dataset contains diverse cell types (e.g., nephrons, endothelial cells, immune cells) with expression profiles significantly different from liver. The model fails to recognize TFs active exclusively in kidney but barely present in the training set, such as certain TFs critical for kidney development but weakly expressed in liver. Consequently, corresponding edge weights in predictions approximate zero and cannot be activated. Future research could explore multi-task learning frameworks to simultaneously learn regulatory features across multiple tissues, or fine-tune models using limited target tissue data (e.g., using partial kidney tissue samples or single-cell data) to narrow this performance gap, which represents a common strategy in transfer learning.

A major advantage of knowledge-driven models is their interpretability, allowing understanding of model parameters from a regulatory interaction perspective. In RNN models, weights associated with TF inputs for hidden state updates can be viewed as the average "importance" of that TF for overall regulation. Although systematic analysis of all weights exceeds the scope of this paper, we can provide several qualitative observations: the model indeed emphasized known major regulatory factors in liver. For instance, TFs such as HNF4A, FOXA2, and CEBPA were assigned higher weights, consistent with their core roles in regulating metabolism-related genes in hepatocytes. These TFs appear frequently in motif networks across multiple genes, and the model's weight enhancement reflects their extensive regulatory functions. In contrast, ubiquitous TFs like SP1, despite widespread DNA binding, were assigned lower effective weights, indicating that while their binding sites are numerous, they are not primary expression regulators. This aligns with existing biological understanding that SP1 may function more as a structural auxiliary factor rather than a primary switch.

Additionally, the model demonstrates capacity to discover potential new regulatory relationships. Since the training objective is expression data, if a TF's expression fluctuations effectively predict target gene expression changes, the corresponding connection weight automatically increases, potentially revealing TF-gene relationships not yet thoroughly investigated. For example, despite limited literature on TBX15, the model frequently utilizes

it to explain expression changes in certain genes, suggesting it may serve important functions in liver. This capability is valuable for generating new biological hypotheses. In practice, such signals can be identified by analyzing trained weight matrices or attention maps.

Notably, negative weights in the model may correspond to inhibitory regulation. Although the input network does not distinguish between activation and inhibition, only by indicating interaction existence, the model can learn negative regulation. After training, only one TF-gene pair with negative weights was observed: EHF and RPS21. While direct evidence that EHF inhibits RPS21 is currently lacking, functional regulation overlap exists between them in cancer. Potential mechanisms include EHF regulating certain miRNAs or upstream TFs, indirectly suppressing RPS21, or EHF activating AKT/ERK pathways, which indirectly affect RPS21 by regulating translation initiation factors and ribosomal proteins. These speculative relationships await further experimental verification.

Our methodological approach has several limitations. Our model shows bias when predicting genes with zero true expression. Using MML activation functions, the model imposes no hard constraints on zero values, causing "off" states to be predicted as extremely low expression. However, functional "off" states do exist in biological systems. This error may lead to misidentification of regulatory activity in certain genes. For critical applications, introducing post-processing threshold mechanisms could be considered, forcing outputs below certain thresholds to zero to more accurately reflect gene on/off states. A more sophisticated approach would employ multi-task learning structures, introducing dual branches in the model: one determining whether genes are "on," and another predicting expression levels, similar to zero-inflated models in statistics. Although not implemented in this study, this approach has potential for handling abundant zero values in single-cell data and tissue-specific genes in bulk data, representing a direction worth exploring in future research.

check this out

This study is based on static data, without explicitly modeling dynamic changes in TF activity over time, signaling pathways, or conditions. In reality, many TFs, though expressed, are not always active (e.g., requiring phosphorylation or ligand binding for activation). We default to using TF mRNA levels as activity proxies, effective in some contexts (like HNF4A in liver) but failing to capture TFs primarily regulated by localization or post-translational modifications (such as NF-κB, p53). Regulatory effects of such TFs may be incorrectly

attributed to other TFs with expression variations. Therefore, constructing more accurate regulatory models may require integrating signaling pathway information or introducing protein-level data, which presents challenges.

Our observation of negative R^2 values in kidney single-cell data suggests that traditional correlation metrics may be insufficient for comprehensive model performance assessment in highly heterogeneous data. Single-cell expression data features widespread zero values and extreme high values; simple comparisons between predicted and observed total values may obscure key trends captured by the model. More appropriate evaluation methods might include predicting average gene expression across all cells or distribution characteristics of expression. To some extent, bulk data evaluation already reflects this average level; another approach is assessing whether the model can predict expression differences between cell types (e.g., identifying which cells have genes "turned on"). These more granular analyses might better reflect model performance. Notably, despite an R^2 value of only 0.30, the Pearson correlation coefficient in liver single-cell data was approximately 0.68, indicating the model indeed captured partial single-cell variation.

Despite these limitations, our modeling framework demonstrates significant effectiveness in explaining gene expression variation and offers stronger interpretability compared to "black box" models due to the incorporation of prior regulatory networks.

6. Conclusions

We constructed a high-resolution human transcription factor-gene regulatory network and developed a recurrent neural network model capable of accurately predicting gene expression states based on transcription factor expression profiles. This model integrates motif-based prior knowledge with transcriptomic data, demonstrating excellent cross-validation performance and external generalization capability, while generating interpretable regulatory weights consistent with established biological features. The framework and findings presented in this study establish a foundation for future integration of additional data modalities and for simulating and understanding regulatory mechanisms across diverse biological contexts. Furthermore, the model facilitates the discovery of novel regulatory relationships through the identification of potentially critical regulatory connections, providing valuable leads for subsequent experimental validation.

7. Points of perspective

1. The framework proposed in this study demonstrates high scalability and can be extended to modeling transcriptional regulation in other tissue types or different species. By incorporating multi-task learning strategies or transfer learning mechanisms, the model can acquire shared regulatory features while capturing tissue-specific or species-specific expression patterns, thereby achieving broad applicability across diverse conditions.

2. The model's performance exhibited notable deterioration when applied to single-cell data, revealing limitations in existing methodologies for addressing cellular heterogeneity. Future research could incorporate zero-inflated modeling, cell type annotation-assisted learning, or architectures combining graph neural networks with variational autoencoders (VAEs) to more accurately parse the sparse, highly variable expression characteristics inherent to single-cell data.

3. This study employed static mRNA expression levels as proxies for TF activity, which precludes the detection of TF activity changes dependent on phosphorylation, nuclear translocation, or ligand activation. Future iterations could integrate phosphoproteomic data, signaling pathway databases, or time-series measurements into the model, thereby enhancing its capacity to model dynamic regulatory processes, particularly under conditions of disease progression, development, or pharmacological stimulation.

4. The TF-based regulatory modeling framework presented herein offers significant potential for clinical transcriptomics applications. In diseases such as cancer, hepatic disorders, or renal pathologies, aberrant activity of specific TFs constitutes a critical driving factor. This model, capable of identifying master regulators and predicting their target gene expression patterns, could be applied to biomarker screening, drug target discovery, and even personalized therapeutic response prediction.

5. Through analysis of the learned edge weights, researchers can generate novel hypotheses regarding TF regulatory functions for subsequent experimental validation. Particularly for regulatory pairs predicted with high confidence but lacking substantial literature evidence, this data-driven mechanism inference capability will accelerate new biological discoveries.

8. Acknowledgements

Click to enter text

9. References

Click to enter text

Appendix (Reflection on ethics, sex and gender perspectives and possible sustainability improvements)

Click to enter text