EMO: Predicting Non-coding Mutation-induced Up- and Down-regulation of Risk Gene Expression using Deep Learning

Zhe Liu[1,#], Yihang Bao[1,#], Weichen Song[3], An Gu[1], Guan Ning Lin[1,2,*]

[1] Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
[2] Shanghai Key Laboratory of Psychotic Disorders, Shanghai, China
[3] Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders (Ministry of Education), Collaborative Innovation Center for Brain Science, Shanghai Jiao Tong University, Shanghai, China

*Authors to whom correspondence should be addressed. (Guan Ning Lin: nickgnlin@sjtu.edu.cn)
[#]These authors contributed equally to this work

Abstract

Motivation

Understanding how changes in non-coding DNAs regulate gene expression remains a formidable challenge with profound implications for advancing human genetics and disease research. Accurate prediction of the up- and down-regulation of gene expression quantitative trait loci (eQTL) can offer a potential solution to expedite the identification of non-coding variants associations with phenotypes. However, despite existing methods for predicting the impact of non-coding mutations on changes in gene expression, the current SOTA tool 'Enformer' still cannot accurately predict the sign of eQTLs. Moreover, the constraints of tissue specificity necessitate the utilization of distinct training models for each particular tissue type within existing methods. This hinders the extension of predictive capacities to the level of single-cell resolution.

Results

In this work, we introduce a novel transformer-based pretrained method, called EMO, to predict the up- and down-regulation of gene expression driven by single non-coding mutations from DNA sequences and ATAC-seq data. It extended the effective prediction range to 1Mbp between the non-coding mutation and the transcription start site (TSS) of the affected gene, with competitive prediction performance across various sequence lengths, outperforming the retrained Enformer structures. We fine-tuned EMO on the eQTLs of two brain tissues to evaluate its robustness through external validation. We also evaluated the transfer ability of EMO into the single-cell resolution by fine-tuning it on six types of immune single-cell eQTL, achieving satisfactory performance in all

cell types (AUC > 0.860). EMO also showed its potential in handling disease-associated eQTLs.

Availability and implementation

The source code is freely available at https://github.com/Liuzhe30/EMO.

1 Introduction

Deciphering gene expression regulation and the effects of genome variation is crucial in human genetics and precision medicine[1-3]. The analyses published to predict the deleteriousness of genetic variation within individual genomes have primarily concentrated on protein-altering variants[4, 5]. However, non-coding variants constitute the overwhelming majority of human genetic variation[6, 7]. The vast noncoding mutation space presents a significant challenge, particularly in understanding how noncoding DNA impacts gene expression in different cell types[8].

Some computational methods have been developed to predict the effect of non-coding mutations on gene expression, such as DeepSea[9], Basenji2[10], Expecto[1] and Enformer[8]. However, these methods can only provide gene expression predictions for regulatory regions within 100kbp upstream or downstream of the non-coding mutation, when the scope of cis-regulation can extend up to 1Mbp[11]. What's more, these methods train models separately for specific tissues, indicating a lack of generalizability, as they can only make predictions for cell types and tissues present in the training data, and cannot be transferred to new cell types or tissues.

Predicting the up- and down-regulation of gene expression quantitative trait loci (eQTL) could accelerate the identification of associations between non-coding variants and phenotypic traits. However, even the state-of-the-art (SOTA) model, Enformer, has been reported to struggle with accurately predicting the sign of eQTL, particularly when the regulatory range exceeds 1kbp[8].

In this study, we introduced EMO, a pre-trained model designed to predict the effect direct of the single non-coding mutations on gene expression from DNA sequences and ATAC-seq data, which can be fine-tuned to new tissues and cell types without the need for training specific models. EMO was designed as a cost-effective segmented model enabling predictions within a regulation range of 1Mbp with interpretability.

We compared the prediction performance of EMO with the retrained and optimized architecture of Enformer and found that EMO's advantage increased as the regulatory range expanded. Additionally, we visualized the embeddings obtained by EMO through representation learning and conducted ablation study to explore the independent contributions of each model component. To demonstrate the capability of EMO for fine-tuning and generalization to new tissues, we collected the eQTLs of two new brain tissues that not present in the training set for external validation. It was observed that

the embeddings output by pre-trained EMO improved the rationality of prediction results compared to the end-to-end predictions. Furthermore, through transfer learning, we evaluated EMO on single-cell eQTLs in six different cell types, achieving satisfactory results across all cell types (AUC > 0.860). Notably, EMO also achieved accurate predictions on two disease-associated eQTLs, suggesting its potential application in deciphering the mechanisms of disease risk SNPs.

2 Methods

2.1 Datasets

We curated cis-eQTL data of 48 tissue types in GTEx v8[12], retaining only those eQTL with a probability of being causal exceeding 0.9 following statistical fine-mapping[13]. Subsequently, we cross-referenced this eQTL dataset with 19 overlapping tissue-specific ATAC-seq data available in EpiMap[14], and discarded the eQTLs with multi-site substitutions for subsequent prediction. The sequences between non-coding mutations and the TSS were extracted from the hg38 human genome. All eQTLs were classified into two categories, 'up-regulation' and 'down-regulation', based on their impact direction on the quantitative expression of the given gene. Due to the potentially biased model resulting from an overrepresentation of entries from heart left ventricle and esophagus mucosa, we applied down-sampling to these two tissues (Figure 1 and Figure S1).

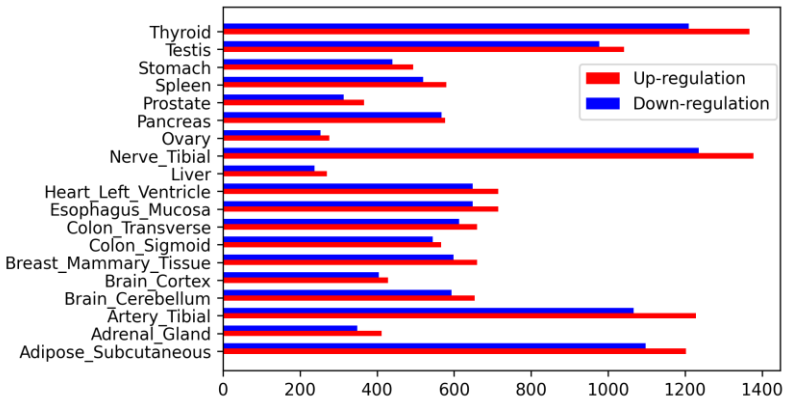

Figure 1. The distribution of positive and negative samples ('up-regulation' and 'down-regulation' eQTLs) in different tissues in the processed dataset.

For each eQTL entry, we utilized the DNA sequence between non-coding mutation and TSS along with its corresponding ATAC-seq as the input features for deep learning. Specifically, in cases where multiple ATAC-seq results corresponded to the same tissue, we conducted a base-wise averaging:

$$CA_i^f = \frac{\sum_{t=1}^{n} CA_i^{T_t}}{n} \tag{1}$$

where CA_i^f represents the final chromatin accessibility of the i-th base in the gene sequence, n signifies the number of included ATAC-seq data for the current

tissue within the dataset, and CA^{Tt} denotes the t-th of the n ATAC-seq data. Due to the broad range of the cis-eQTLs' impact, extending from within 1,000 bp to as far as 1 million bp, blind data completeness for alignment purposes could potentially lead to unnecessary computational waste. Thus, based on the varying magnitudes of the extracted DNA sequence lengths, we partitioned the dataset into four distinct segments: the "small" dataset (1-1,000 bp), the "middle" dataset (1,001-10,000 bp), the "large" dataset (10,001-100,000 bp), and the "huge" dataset (100,001-1,000,000 bp). The training, validation and test dataset were randomly distributed based on an 8:1:1 ratio (Table S1, Table S2, Figure S2).

We collected eQTLs of two new brain tissues, hippocampus and spinal cord, from MetaBrain[15] for external validation (Table S7), and their corresponding tissue-specific ATAC-seq data was downloaded from EpiMap, undergoing the same data preprocessing procedure. Additionally, to validate the transferability of EMO at the single-cell resolution, we acquired single-cell eQTLs from six cell types from the OneK1K cohort[16] and corresponding cell line ATAC-seq data from EpiMap (Table S10).

2.2 Model architecture and prediction pipeline

We formulate the sign prediction of eQTLs as a 2-class classification problem, where each non-coding mutation that affected the gene expression may be assigned as either 'up-regulation' and 'down-regulation'. EMO takes the One-Hot coding of the DNA sequence between the non-coding mutation and TSS, 51bp of the One-Hot coding of the DNA sequence of the reference genome centered on the mutation, 51bp of the One-Hot coding of the DNA sequence of the mutated sequence centered on the mutation, and the corresponding chromatin accessibility of the DNA sequence as input. The model's output represents the predicted probabilities of a non-coding mutation affecting gene expression in an up-regulated manner.

EMO consists of five main stages (Figure 2): (i) encoding the DNA sequence between the non-coding mutation and TSS as 'Between Branch', (ii) utilizing bidirectional Gated Recurrent Units (BiGRUs), aligning sequences of different lengths based on distinct-scale pooling strategies, (iii) applying rotary position embedding to the aligned features, subsequently inputting them into sparse multi-head attention layers, (iv) encoding the DNA sequence with a length of 51bp centered on the non-coding mutation site before and after the mutation, and inputting them into convolutional neural networks (CNN) as 'Mutation Branch', (v) merging the feature representation from both the 'Between Branch' and the 'Mutation Branch' and utilizing fully-connected layers and the SoftMax function[17] for the final prediction, which are described in Sections 2.3, 2.4 and 2.5.

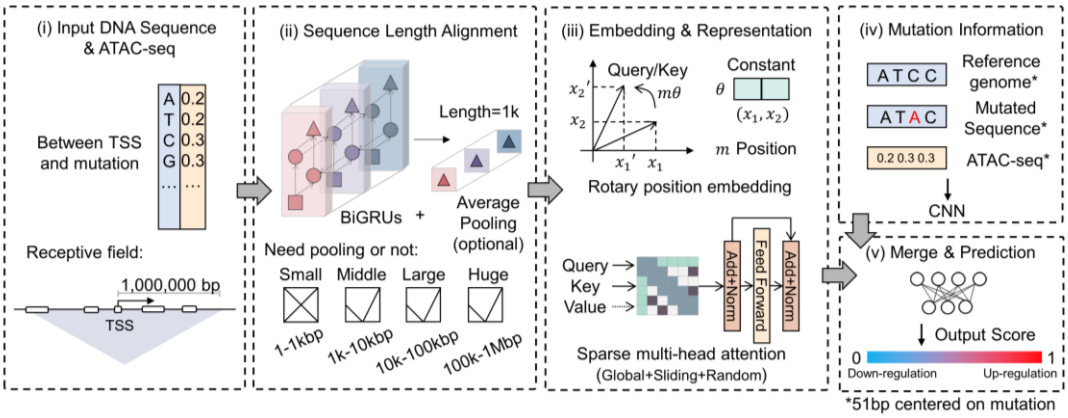

Figure 2. An overview of the EMO pipeline for non-coding mutation effect on eQTL sign prediction. The input of the model contains the 'Between Branch' and the 'Mutation Branch', One-Hot coding of the DNA sequence between the non-coding mutation and TSS, and the corresponding chromatin accessibility of the DNA sequence. The output is the predicted probabilities of a non-coding mutation affecting gene expression in an up-regulated manner.

2.3 Alignment of DNA sequence with pooling

Since cis-eQTLs are SNPs within 1Mbp of a gene's transcription start[11], aligning all DNA sequences to 1Mbp would result in computational waste for samples with medium to short DNA sequences. Therefore, we padded the input features with zeros based on the original length of DNA sequences: the sequences within the range of 1bp to 1kbp were padded to 1kbp, those within 1kbp to 10kbp were padded to 10kbp, and those within 10kbp to 100kbp were padded to 100kbp. Sequences longer than 100kbp were padded to 1Mbp.

After BiGRUs, we also conducted average pooling[18] for the sequences with lengths exceeding 1kbp:

$$f_{avg}(x) = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2}$$

where x is a vector consist of activation values from a rectangular area of N vectors in the dimension of the input features, keeping the embedding feature dimensions constant, we compressed the dimensions of the input sequences. The pooling size varied according to the magnitude of the length of the DNA sequences:

$$N = \begin{cases} 10, & L = 10kbp \\ 100, & L = 100kbp \\ 1{,}000, & L = 1Mbp \end{cases} \tag{3}$$

where L is the length of the input sequence after padding. Ultimately, the output feature of all DNA sequences was aligned to 1000 dimensions.

2.4 Representation learning with sparse Transformer

For the aligned 1000-dimensional output, we employed rotary position embedding[19] to indicate the positional information:

$$f_{\{q,k\}}(x_m, m) = R^d_{\theta,m} W_{\{q,k\}} x_m \tag{4}$$

where $x_m \in \mathbb{R}^d$ is the d-dimension feature embedding vector of the m-th token without position information, and $R^d_{\theta,m}$ is the rotary matrix with pre-defined parameters $\theta = \{\theta_i = 10000^{-2(i-1)/d}, i \in [1,2,\dots,d/2]\}$.

Then, we implemented sparse self-attention layers based on BigBird[20], which contains the global, sliding and random multi-head self-attention for representation learning. The number of attention heads for each self-attention layer was set to 8.

2.5 Convolution and merge of mutation information

To preserve information from the reference genome near the mutation site and enhance the learning of mutation information with the network, we extracted reference genome and post-mutation DNA sequences with a total length of 51bp, centered around the mutation site, along with their corresponding ATAC-seq data. Then, this portion of the input as 'Mutation Branch' was fed into a convolutional module consisting of 2 CNN layers.

Then, we concatenated the embeddings from the 'Between Branch' and the 'Mutation Branch' and fed them into three fully-connected network layers with decreasing numbers of artificial neurons. For the final output, we utilized the SoftMax function[17] , resulting in the prediction outcome of EMO.

2.6 Evaluation metrics

As our study focuses on two-class classification for clinical drug response prediction, we evaluated the performance of our predictor using the following metrics[21]: accuracy, recall, precision, F1-score, and AUC, which were defined as:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$F1 - score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{8}$$

where TN, TP, FN, and FP represent the number of correctly predicted non-sensitive mutations, the number of correctly predicted sensitive mutations, the number of falsely predicted non-sensitive mutations, and the number of falsely predicted

sensitive mutations, respectively. The AUC is computed by using the trapezoidal rule to approximate the region under the receiver operating characteristic curve.

3 Results

3.1 Model performance in short- and long-range prediction

We trained EMO as four sub-models, each dedicated to handling non-coding mutations within different impact ranges (length of DNA sequence). These sub-models are referred to as 'EMO-small' (1bp-1kbp), 'EMO-middle' (1kbp-10kbp), 'EMO-large' (10kbp-100kbp), 'EMO-huge' (100kbp-1Mbp). The performance of these models was evaluated on the independent test sets, as illustrated in Figure 3. Despite a gradual decrease in the models' capability to identify non-coding mutations affecting as 'expression down-regulated' with increasing sequence lengths, all sub-models exhibited an overall predictive AUC of 0.7 or higher.

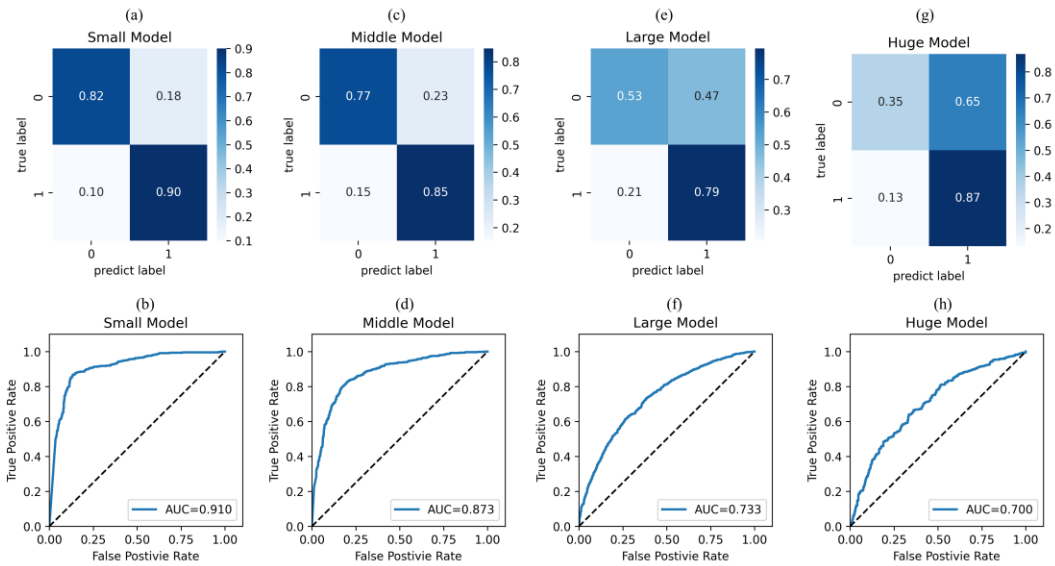

Figure 3. Model performance of different sub-models of EMO of different sequence lengths on the independent test set.

Considering that Enformer has not publicly released the pre-trained tissue-specific sub-models, we retrained the network architecture of Enformer in our datasets and optimized it to achieve overall prediction performance better than that reported in the original paper[8] to make it support tissue-specific eQTL sign prediction, and compared its performance with EMO.

As shown in Table 1, both EMO and Enformer demonstrate relatively accurate prediction performance (EMO AUC=0.910, Enformer AUC=0.894) when the DNA sequence length ranges from 1 to 1kbp, which is consistent with the previous reported results[8]. However, as the length of DNA sequence increases, EMO exhibits a superior

prediction capacity, potentially attributed the adverse effects of an excessive number of CNN layers in Enformer's architecture when handling longer DNA sequences.

Table 1. Model performance and comparison of EMO and retrained Enformer.

Sequence length	Model	ACC	Precision	Recall	F1-score	AUC
1-1k	EMO	**0.868**	**0.870**	**0.900**	**0.885**	**0.910**
	Enformer	0.833	0.867	0.831	0.849	0.894
1k-10k	EMO	**0.819**	**0.810**	**0.845**	**0.827**	**0.873**
	Enformer	0.629	0.674	0.530	0.594	0.674
10k-100k	EMO	**0.667**	**0.643**	0.794	**0.711**	**0.733**
	Enformer	0.516	0.516	**1.000**	0.681	0.524
100k-1M	EMO	**0.648**	**0.643**	**0.868**	**0.739**	**0.700**
	Enformer	-	-	-	-	-

Numbers in bold indicate the best experimental results.
'-' indicates that the length of DNA sequences exceeds the prediction capability of the model and the result cannot be obtained.

3.2 Representation visualization and ablation study

To investigate the representation ability of EMO and its capacity to extract high-level features the input data, we used t-SNE[22] to visualize the initial features of the input branches and the final features prior to the output layer on the entire training dataset. As shown in Figure 4(a)-(h), no distinct or meaningful clusters were evident in the input features. However, after training, samples representing the two distinct categories exhibited a noticeable separation trend within the embedding space (Figure 4(i)-(l)), which became more pronounced in the case of short to medium DNA sequences, indicating that EMO could identify the proper classification hyperplane for the training dataset.

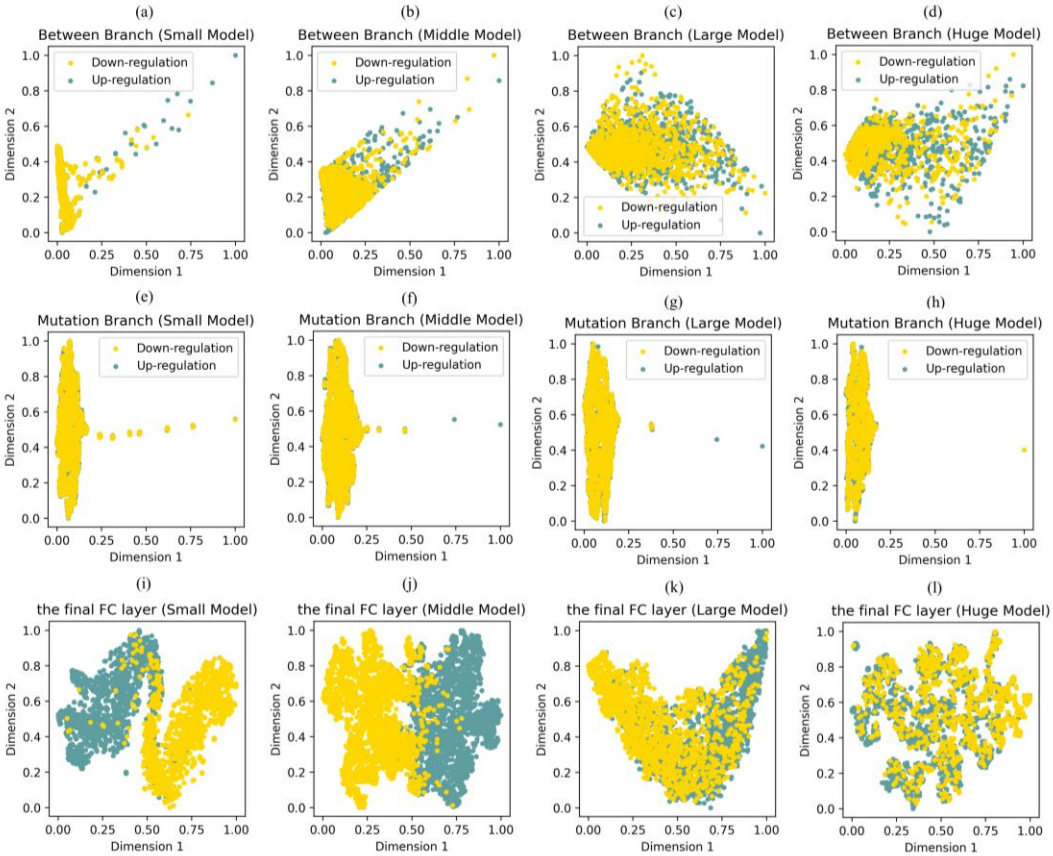

Figure 4. Representation visualization of EMO with input feature and the learned embedding before the output layer.

As the model comprised several processing modules, to explore the independent contributions of each module to the model's performance, we conducted an ablation experiment (Table 2). It was observed that the DNA sequence is undoubtedly the most crucial factor, while the ATAC-seq data, capable of reflecting tissue-specific information due to the tissue-specificity of eQTLs, is equally indispensable. Regarding the model architecture, we assessed the necessity of the architecture by replacing the original network layers with fully connected layers, and found that the BiGRUs for capturing local information and the sparse self-attention layers for extracting long-range dependencies both made distinct individual contributions. \

Table 2. Ablation study on the 'EMO-Middle' model.

Component	Option	ACC	Precision	Recall	F1-score	AUC	Parameters
Input feature	Without ATAC-seq	0.762	0.793	0.723	0.756	0.838	2,568,722
	Without DNA sequence	0.589	0.623	0.494	0.551	0.619	2,494,994
Local feature block	Without BiGRU*	0.745	0.777	0.701	0.737	0.811	2,486,162
Long-range	Without	0.724	0.724	0.744	0.734	0.804	1,947,378

block	attention layers*						
	EMO	**0.819**	**0.810**	**0.845**	**0.827**	**0.873**	2,594,298

Numbers in bold indicate the best experimental results.
'*' indicates that the component of EMO was replaced by fully-connected layers, maintaining the same output dimensions.

3.3 Attention interpretability

Regulatory features like transcription factors and chromatin markers play vital roles in gene expression regulation and has been reported to be effective in associating eQTLs with target genes[23, 24]. In other words, in the sequence between non-coding mutations and the TSS, regulatory feature regions will provide more information for eQTLs sign prediction compared to other regions. Therefore, in order to investigate whether our model placed more emphasis on the information from such regulatory feature regions to enhance its accuracy, we map regulatory features to our input sequence and observe whether the attention weights of our model are significantly enriched in these regions.

We first obtained enhancer, TF binding, promotor, CTCF binding site and open chromatin annotations from the Ensembl database[25] (version 2023-7-28) and mapped them to our training set input sequences. Then, we expanded the output attention weight vector to its original input length and separately aggregated the weights of regulatory feature regions and other regions. Due to the average length of regulatory feature regions being 950, sequences that are too short would be excessively annotated, and sequences that are too long would introduce larger errors due to long pooling step. Here, we only conducted this attention interpretability part on sequences of middle and large lengths. The Mann-Whitney one-tailed test[26] results showed that, regulatory feature regions had significantly greater attention weights than other regions in 75% of the heads (Table 3). This indicates that our model tends to focus more on the regulatory feature regions within the sequences, which is consistent with the previously mentioned relevance of regulatory features to the eQTL sign prediction task.

We also visualized two examples to further support the above conclusion (Figure5). The eQTL for the small model is located at position 324,170 on chr11, from A to G, with the gene's TSS downstream at position 5,530. And the eQTL for the large model is located at position 15,194,525 on chr10, from A to G, with the gene's TSS downstream at position 25,832. By comparing the attention input and attention output, we can observe that the model amplifies its focus on the regulatory feature region through the attention module.

Table 3. Mann-Whitney one-tailed test p-values for attention weights in regulatory feature regions vs. other regions

Model	Head1	Head2	Head3	Head4
Middle	5.42e-09	2.40e-43	1	1.27e-37
Large	1	0	0	0

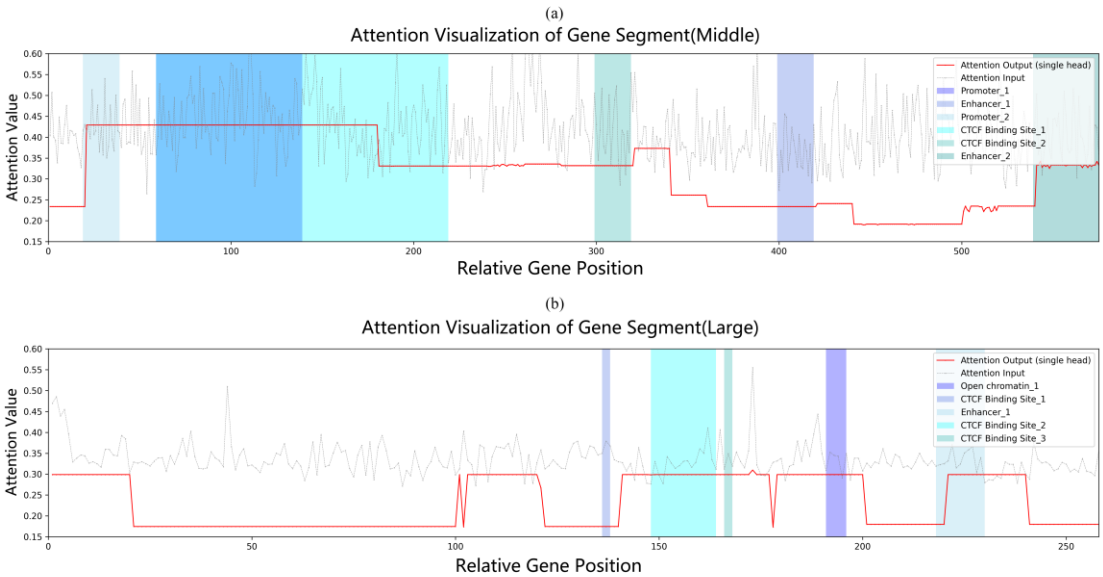

Figure 5. Visualization of attention IO and regulatory feature regions for two examples. gray lines represent attention input, red lines represent attention output from head2. colored segments denote annotated regulatory feature regions.

3.4 Tissue-specific fine-tuning on external brain eQTLs

As a tissue-agnostic universal pre-trained mode, EMO can seamlessly leverage tissue-specific ATAC-seq data and fine-tune it to various tissues, and even migrate to single-cell resolution, by learning the association between tissue-specific impact patterns of non-coding mutation and tissue ATAC-seq data.

We fine-tuned the EMO model on tissues with a relatively higher number or entries in the training set. Comparing these results with those of the end-to-end trained EMO model on tissue eQTLs separately, we observed that the fine-tuned predictions significantly outperformed the end-to-end predictions (Figure S3 and Table S4-6).

Subsequently, we obtained the non-coding mutations from two brain tissues, hippocampus and spinal cord, from MetaBrain[15], which were not included in the EMO training set. As depicted in Figure 6, the fine-tuned EMO exhibited more reasonable predictions compared to the end-to-end predictions. This finding suggests that the prior information contained in the embedding representation obtained from the pretrained EMO contributes to better performance in predicting the up- and down-

regulation of non-coding mutation effects for tissues with small sample sizes.

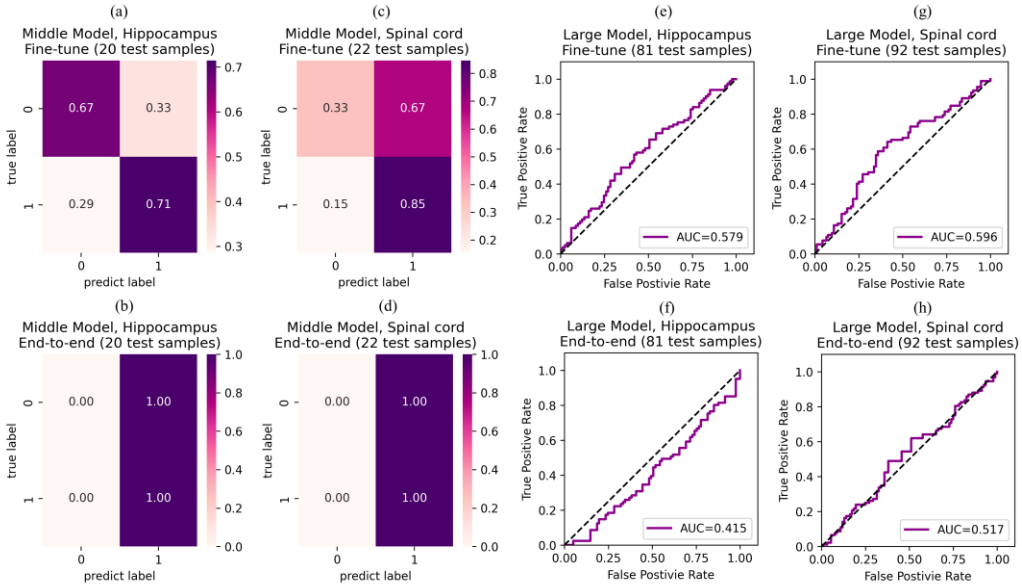

Figure 6. The prediction results of the fine-tuning and the end-to-end EMO model on eQTLs of two new tissues, hippocampus and spinal cord.

3.5 Transfer EMO to single-cell eQTLs

One of the core advantages of EMO is the ability to transfer pre-trained embedding to predict the up- and down-regulation of single-cell non-coding mutations through transfer learning. We collected the single-cell eQTL data of six immune cell types from the OneK1K cohort[16] and performed fine-tuning using EMO individually (Table 4 and Figure S5).

It is observed that EMO demonstrated satisfactory performance across all cell types (AUC > 0.860), indicating the generalization potential of EMO's pre-trained embedding representation in predicting the up- and down-regulation of non-coding mutation effects at single-cell level, offering new insights into the interpretation of cellular-scale microenvironments.

Table 4. The results of the transfer fine-tuned 'EMO-large' model on single-cell eQTLs of six cell types.

Cell type	ACC	Precision	Recall	F1-score	AUC
B naïve	0.815	0.882	0.600	0.714	0.861
CD4 memory	0.884	1.000	0.853	0.921	0.923
CD4 naïve	0.865	0.864	0.826	0.844	0.948
CD8 memory	0.756	0.714	0.870	0.784	0.862
Dendritic cell	0.818	0.893	0.893	0.893	0.919
Natural killer cell	0.823	0.913	0.677	0.778	0.911

3.6 Analysis of disease-associated eQTL cases

To further investigate the performance of EMO on non-coding mutations associated with diseases, we collected two reported disease-associated eQTLs as exploratory cases[15]. The A allele of variant rs4698412 was reported increased the expression of *CD38* and associated with increased risk for Parkinson's disease (PD), and rs1902660 was analyzed as an eQTL with higher expression of *TSPAN14* as putatively causal for Alzheimer's Disease (AD).

We matched these two eQTLs with ATAC-seq data of cerebellum[27, 28] and used EMO for prediction. As shown in Table 5, EMO correctly identified the direction of the effect of these two disease-associated non-coding mutations on gene expression, demonstrating its potential in deciphering the mechanisms of action at disease risk SNPs.

Table 5. Evaluation on two disease-associated SNPs.

Disease	RSID	Gene	True label	EMO prediction result	Prediction score
Alzheimer's Disease	rs1902660	*TSPAN14*	Up-regulation	Up-regulation	0.937
Parkinson's Disease	rs4698412	*CD38*	Up-regulation	Up-regulation	0.690

4 Discussion

An enduring challenge in regulatory genomics has been the accurate prediction of gene expression based solely on DNA sequence, and precise prediction of the sign of eQTL can accelerate the discovery of associations between non-coding variants and phenotypes. In this study, we introduced EMO, a pre-trained model for predicting the up- and down-regulation of single non-coding mutations on gene expression, which can be fine-tuned to any tissues or cell types when the ATAC-seq data was provided. We conducted a cost-effective exploration of long-range DNA sequence regulatory patterns and provided interpretability. Our work offers a novel perspective on the impact of non-coding risk SNP on the transcriptome, pinpointing this interpretation to single-cell resolution. Additionally, we validated EMO on multiple external datasets and illustrated its prediction potential on disease-associated risk SNPs.

However, EMO exhibits some limitations: to distinguish between up-regulated and down-regulated non-coding mutations, EMO tends to provide bolder rather than conservative predictions. This might lead to severe misjudgments and the inability to recognize the mutations with 'no effect', making it challenging for EMO to handle large-scale screening tasks. It is recommended to consider EMO as a supplement method under the premise of known risk SNPs. Furthermore, there is still room for improvement of EMO's prediction performance for medium to long DNA sequences.

In the future, we will continue to optimize EMO, enabling it to rapidly annotate genes

that may be affected by specific risk non-coding mutations on a large scale and provide the direction of such effects. With the advancement of large language pre-trained models, we also consider incorporating semantic embedding of DNA sequences for optimizing EMO.

Conflict of interest

The authors declare that they have no competing interests.

Funding

This work was supported by grants from STI 2030—Major Projects (no. 2022ZD0209100), the National Natural Science Foundation of China (nos. 81971292 and 82150610506), the Natural Science Foundation of Shanghai (no. 21ZR1428600), the Medical-Engineering Cross Foundation of Shanghai Jiao Tong University (nos. YG2022ZD026 and YG2023ZD27), and SJTU Trans-med Awards Research (no. 20220103)

References

1. Zhou, J., et al., *Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk.* Nature genetics, 2018. **50**(8): p. 1171-1179.
2. Ashley, E.A., *Towards precision medicine.* Nature Reviews Genetics, 2016. **17**(9): p. 507-522.
3. Mohammadi Pejman 5 6 Park YoSon 11 Parsana Princy 12 Segrè Ayellet V. 1 Strober Benjamin J. 9 Zappala Zachary 7 8, G.C.L.a.A.F.B.A.A.C.S.E.D.J.R.H.Y.J.B., et al., *Genetic effects on gene expression across human tissues.* Nature, 2017. **550**(7675): p. 204-213.
4. Cooper, G.M. and J. Shendure, *Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data.* Nature Reviews Genetics, 2011. **12**(9): p. 628-640.
5. Lek, M., et al., *Analysis of protein-coding genetic variation in 60,706 humans.* Nature, 2016. **536**(7616): p. 285-291.
6. Consortium, G.P., *A map of human genome variation from population scale sequencing.* Nature, 2010. **467**(7319): p. 1061.
7. Khurana, E., et al., *Role of non-coding sequence variants in cancer.* Nature Reviews Genetics, 2016. **17**(2): p. 93-108.
8. Avsec, Ž., et al., *Effective gene expression prediction from sequence by integrating long-range interactions.* Nature methods, 2021. **18**(10): p. 1196-1203.
9. Zhou, J. and O.G. Troyanskaya, *Predicting effects of noncoding variants with deep learning–based sequence model.* Nature methods, 2015. **12**(10): p. 931-934.
10. Kelley, D.R., *Cross-species regulatory sequence activity prediction.* PLoS computational

biology, 2020. **16**(7): p. e1008050.
11. Gaynor, S.M., et al., *Connectivity in eQTL networks dictates reproducibility and genomic properties.* Cell Rep Methods, 2022. **2**(5): p. 100218.
12. Consortium, G., *The GTEx Consortium atlas of genetic regulatory effects across human tissues.* Science, 2020. **369**(6509): p. 1318-1330.
13. Wang, Q.S., et al., *Leveraging supervised learning for functionally informed fine-mapping of cis-eQTLs identifies an additional 20,913 putative causal eQTLs.* Nature Communications, 2021. **12**(1): p. 3394.
14. Boix, C.A., et al., *Regulatory genomic circuitry of human disease loci by integrative epigenomics.* Nature, 2021. **590**(7845): p. 300-307.
15. de Klein, N., et al., *Brain expression quantitative trait locus and network analyses reveal downstream effects and putative drivers for brain-related diseases.* Nature Genetics, 2023. **55**(3): p. 377-388.
16. Yazar, S., et al., *Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease.* Science, 2022. **376**(6589): p. eabf3041.
17. Gibbs, J.W., *Elementary principles in statistical mechanics: developed with especial reference to the rational foundations of thermodynamics*. 1902: C. Scribner's sons.
18. Sharma, S. and D.R. Mehra, *Implications of Pooling Strategies in Convolutional Neural Networks: A Deep Insight.* Foundations of Computing and Decision Sciences, 2019. **44**: p. 303-330.
19. Su, J., et al., *Roformer: Enhanced transformer with rotary position embedding.* arXiv preprint arXiv:2104.09864, 2021.
20. Zaheer, M., et al., *Big bird: Transformers for longer sequences.* Advances in neural information processing systems, 2020. **33**: p. 17283-17297.
21. Salavati, H., et al., *Drug transport modeling in solid tumors: A computational exploration of spatial heterogeneity of biophysical properties.* Computers in Biology and Medicine, 2023. **163**.
22. van der Maaten, L. and G. Hinton, *Viualizing data using t-SNE.* Journal of Machine Learning Research, 2008. **9**: p. 2579-2605.
23. Wang, D., A. Rendon, and L. Wernisch, *Transcription factor and chromatin features predict genes associated with eQTLs.* Nucleic Acids Research, 2012. **41**(3): p. 1450-1463.
24. Flynn, E.D., et al., *Transcription factor regulation of eQTL activity across individuals and tissues.* Cold Spring Harbor Laboratory, 2021.
25. Cunningham, F., et al., *Ensembl 2022.* Nucleic Acids Research, 2021. **50**(D1): p. D988-D995.
26. Mcknight, P.E. and J. Najab, *Mann-Whitney U Test*. 2010: The Corsini Encyclopedia of Psychology.
27. Wu, T. and M. Hallett, *The cerebellum in Parkinson's disease.* Brain, 2013. **136**(Pt 3): p. 696-709.
28. Jacobs, H.I.L., et al., *The cerebellum in Alzheimer's disease: evaluating its role in cognitive decline.* Brain, 2018. **141**(1): p. 37-47.