scientific reports

Check for updates

OPEN

IST: an ontology-guided attention-based autoencoder for interpretable analysis of single-cell transcriptomic data

Tasbiraha Athaya[1], Xiaoman Li[2]✉ & Haiyan Hu[1]✉

Extracting biologically interpretable insights from single-cell RNA sequencing (scRNA-seq) data remains a major challenge. While deep learning approaches have demonstrated strong predictive performance for dimensionality reduction and representation learning, their latent representations often lack clear biological meaning, and prior biological knowledge is frequently incorporated only weakly or inconsistently. As a result, learned features may not faithfully reflect known gene-function relationships, limiting biological interpretability and downstream insight. We present IST (Interpretation of Single-cell Transcriptomic data), a novel attention-based framework designed to reliably integrate biological prior knowledge into the analysis of single-cell transcriptomic data. IST incorporates gene ontology structure directly into the model architecture and employs attention mechanisms to capture context-dependent gene activity. A tailored correlation-based loss enforces alignment between learned gene activities and known gene-function relationships, ensuring that biological priors are faithfully reflected in the learned representations rather than acting as superficial regularization. We evaluate IST on three scRNA-seq datasets. Across all datasets, IST identifies condition-specific gene activity patterns, uncovers novel relationships among biological processes, and infers previously unannotated gene functions supported by existing literature. Compared with state-of-the-art methods, IST provides more interpretable and biologically grounded representations of gene functions and activities, while maintaining competitive predictive performance. The datasets and tool are available at https://doi.org/10.6084/m9.figshare.30811334.

Deep learning (DL) approaches have become powerful tools for analyzing high-dimensional single-cell RNA sequencing (scRNA-seq) data[1–3]. These methods enable dimensionality reduction[4], data integration[5], denoising[6], clustering[7], drug treatment effect prediction[8], and perturbation modeling[9], while capturing complex, non-linear relationships in data that linear methods such as principal component analysis or matrix factorization cannot readily identify[10]. Among them, autoencoders (AEs) and variational autoencoders (VAEs) have become widely used for learning compact and biologically meaningful representations[11,12]. However, they remain limited by their opaque latent spaces, which hinder interpretation in terms of underlying biological mechanisms[11].

To improve interpretability, biologically informed DL frameworks increasingly integrate biological priors directly into model architectures. The interpretable DL approaches pursue two main paths toward model interpretability. One is the post-hoc methods such as LIME[13] or SHAP[14] that approximate feature contributions after training. The second, and typically more effective, strategy incorporates biological priors directly into the network structure, thereby constraining the latent space or decoder to align with meaningful biological entities. Examples of the latter include expiMAP[15] and VEGA[16], which incorporate gene programs into their decoders to align latent dimensions with biological processes. Although effective, their shallow decoder structures restrict modeling of hierarchical or context-dependent biological relationships. To address the need for hierarchical modeling, several ontology, pathway, or transcription factor-guided approaches have been developed. Models like DCell[17], P-net[18], OntoVAE[12], DC[19] and INSIGHT[8] structure their network layers according to biological hierarchical information like hierarchical cell subsystems, pathways, features from patients, biological processes like Gene Ontology (GO), etc. Although the models incorporate biological information, they still cannot disentangle covariate effects (cell types, timepoint, treatment effect, etc.) in their latent space. COBRA[20]

[1]Department of Computer Science, University of Central Florida, Orlando, FL 32816, USA. [2]Burnett School of Biomedical Sciences, College of Medicine, University of Central Florida, Orlando, FL 32816, USA. ✉email: xiaoman@mail.ucf.edu; haiyan.hu@ucf.edu

addresses this by incorporating adversarial disentanglement to separate known covariates, but relies on labeled metadata and sensitive hyperparameter tuning, limiting scalability.

In parallel, transformer-based architectures[21], have recently emerged as powerful tools in scRNA-seq analysis tasks like multi-omics data integration[22,23], data augmentation[24], cell-type annotation[25–27], etc. By leveraging self-attention mechanisms, transformers can capture long-range dependencies and complex, non-linear relationships among genes and cells that conventional architectures often fail to model effectively. Unlike traditional autoencoders or VAEs, which rely on local latent structures, transformers allow the model to weigh the relative importance of features under different conditions, enabling more nuanced representation learning. Although transformer models also have the black-box problem because of its complex structure, they are relatively more interpretable via attention weights. This has motivated the development of biologically informed DL transformer models designed to bridge the gap between predictive power and mechanistic insight. However, existing transformer frameworks like Pathformer[23] lack hierarchical biological grounding, reducing their direct interpretability in terms of functional pathways, GO terms, etc[28].

To address these limitations, we propose IST (*I*nterpretation of *S*ingle-cell *T*ranscriptomic Data), a novel AE framework that integrates attention mechanisms into ontology-informed decoders and employs a tailored correlation-based loss to align learned representations with biological prior knowledge. Existing ontology-informed models, such as OntoVAE and DCell, encode biological hierarchies through masked linear layers. However, their representations remain static and constrained by predefined parent-child relationships. Consequently, interactions between biological processes are limited to the fixed ontology structure and cannot adapt to context-specific dependencies across conditions. In contrast, IST augments the ontology-guided architecture with transformer-based self-attention applied within each ontology layer. This design enables dynamic, context-dependent modeling of relationships among GO terms, allowing the model to capture cross-process and non-local functional dependencies that are not explicitly encoded in the ontology. As a result, IST extends representational capacity beyond fixed hierarchical propagation to adaptive interaction modeling within biologically constrained architectures. Furthermore, the proposed loss function enhances both biological fidelity and latent-space interpretability by explicitly encouraging consistency between learned features and known biological structure, rather than passively incorporating prior knowledge. Together, these contributions: (i) integrating self-attention within ontology layers, (ii) combining hierarchical masking with dynamic cross-process interaction modeling, and (iii) introducing biologically aligned supervision through the loss function, enable IST to capture both hierarchical and context-dependent biological relationships while preserving interpretability. We evaluated IST on three scRNA-seq datasets: mouse skeletal muscle regeneration across four post-injury time points[29], interferon-β–treated human scRNA-seq data of peripheral blood mononuclear cells (PBMC)[30], and mouse embryonic fibroblasts stimulated with interferon-β for 0, 1, or 6 h (MEFs). Across all datasets, IST consistently identified context-specific GO activity patterns and highlighted key biological processes.

Materials and methods

Dataset pre-processing

GO data was obtained from geneontology.org (format-version: 1.2), along with the corresponding Gene Annotation file specific to mouse and human (mgi.gaf, gao_human.gaf, version 2.2). Although we focused exclusively on the "biological_process" GO here, our approach can be readily applied to other GOs and biological priors. We retained only the "is_a" relationships to construct the hierarchical structure. To refine the ontology for modeling purposes, we excluded GO terms associated with <30 or >1000 genes and properly transferred their GOs to remaining GO terms. This filtering resulted in an ontology with 2,978 GO terms and 11,467 genes for mouse skeletal muscle, 3,042 GO terms and 12,355 genes for human PBMC, and 3,300 GO terms and 21,432 genes for mouse MEFs.

We downloaded gene expression data of mouse regenerating muscle tissue from GSE143437, mouse MEFs from GSE160764, and human PBMC from GSE96583. GSE143437 contains scRNA-seq data of mouse muscle at four timepoints of post notexin-injury: day 0, day 2, day 5, and day 7[31]. Day 0 (uninjured) represents the quiescent baseline, where muscle stem cells preserve transcriptomic stability. By Day 2 (early activation and proliferation), inflammatory and cytokine signaling dominate as immune cells infiltrate and muscle progenitors begin to proliferate. Day 5 (growth and differentiation) marks the onset of anabolic and myogenic programs, driving cell growth, biosynthesis, and differentiation. Finally, Day 7 (remodeling and maturation) features structural reorganization, angiogenesis, and extracellular remodeling, leading to the restoration of functional muscle architecture. GSE96583 contains scRNA-seq data from 2 controls and 8 PBMC treated lupus patients using interferon (IFN)-ß. GSE160764 contains scRNA-seq data of mouse MEFs, which has been stimulated using interferon (IFN)-ß for 0 h, 1 h, and 6 h. The genes that have greater than 0 expression in at least 1% of the total number of samples were kept for further analysis. Next, an annotation matrix was made using the genes and the GAF file. The gene that is associated with the GO term was given value 1 and otherwise value 0. Only GO terms that are present in our pruned ontology are used in the matrix. The gene expression matrix is used as the input of the proposed model and the gene-GO matrix is used in the last decoder layer. The data was randomly divided into 70%, 15%, and 15% at each stage for training, validation, and testing, respectively.

Model architecture

Encoder structure

The encoder of IST transforms the high-dimensional gene expression input into a lower-dimensional latent vector space (Fig. 1A). It is built as a sequence of repeating blocks, with each block containing a linear layer for affine transformation and a normalization layer for stabilizing the learning process by normalizing the outputs of the linear layer. A leaky ReLU activation function with a negative slope of 0.01 to introduce non-linearity while preventing the dying ReLU problem. The input vector (e.g., of dimension 11467 in mouse skeletal muscle) first

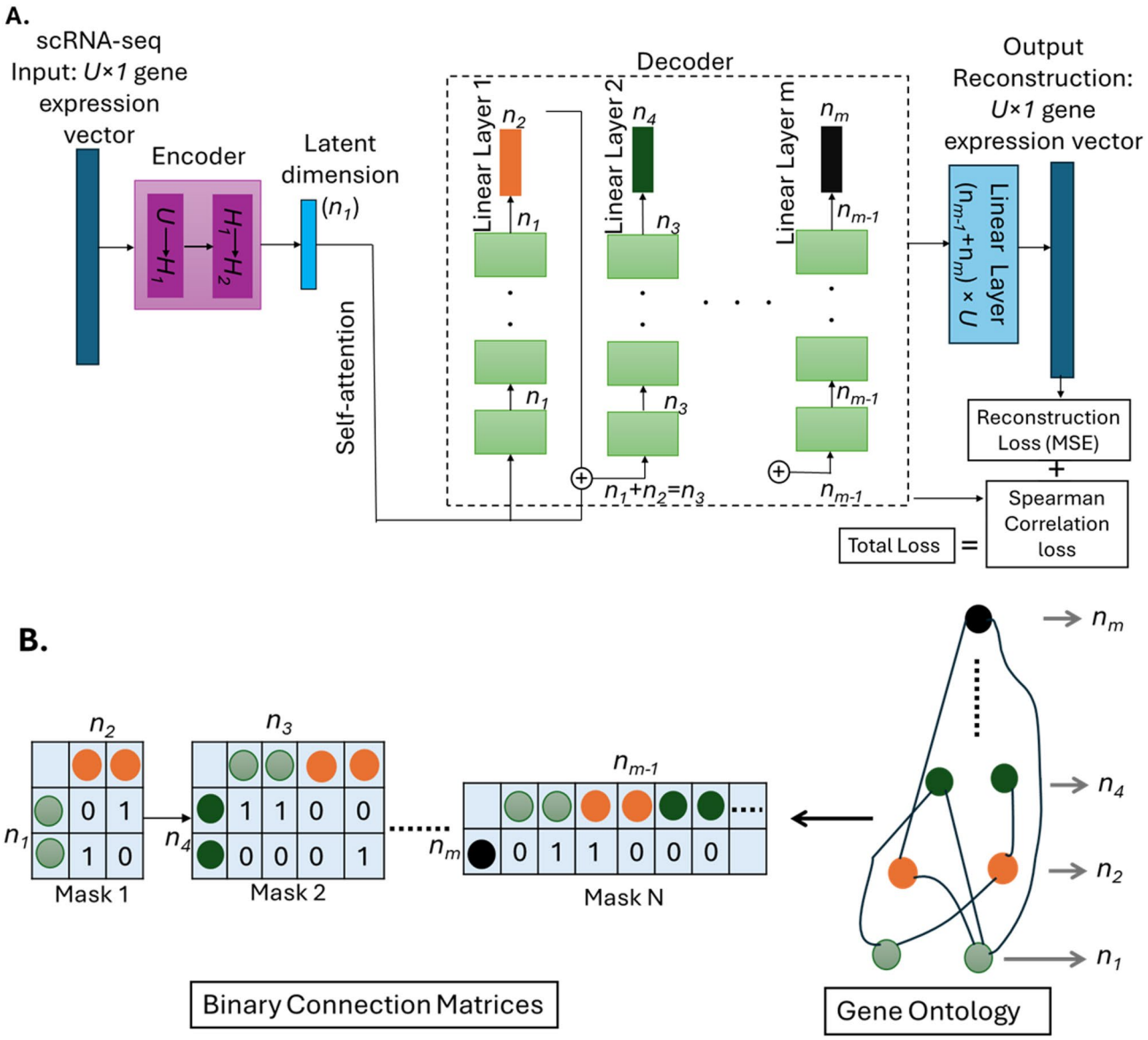

Fig. 1. The overall architecture of proposed IST model. (**A**) The decoder represents the skip connection modeling through concatenations. Binary masks indicate the connection between transformer layers and decoder linear layers (orange, green, black), and output layer (blue). At each step, the previous layer is concatenated to the current one. Mask 1 is modeling the connections between the output of transformer layer 1 and the orange layer, the orange layer is then concatenated to the output of transformer layer 1, so that connections of both layers to the green layer can be modeled by mask 2; (**B**) An example gene ontology (GO) graph and the binary connection matrix of the graph for each layer of decoder.

passes through a Linear layer that maps it to the first hidden dimension (e.g., 512). Then, a single Linear layer projects the output to the final latent dimension (e.g., n_1). The resulting vector, meaning the latent dimension (n_1), is the compressed representation of the expression of a high-dimensional input cell.

Decoder with attention mechanism structure

One novelty of IST is in its decoder (Fig. 1A). Different from the original AE decoder that reconstructs gene expression profile from the latent space representation, the attention-based decoder in IST is designed to reconstruct GO activities from the latent vector and then reconstruct the original gene expression data from the inferred GO activity tree.

The decoder in IST is built to flexibly integrate any Directed Acyclic Graph (DAG), such as a biological process ontology, through the use of linear layers placed after each transformer layer. For this study, we specifically incorporated the biological process GO DAG. DAGs are ideal for this kind of integration due to their hierarchical organization, lack of cycles, and their allowance for nodes to have multiple parent terms, unlike a strict tree structure. We distinguish between the concepts of tier and hierarchical distance: a tier represents the shortest

path from a node to the root, whereas the hierarchical distance of a node refers to the longest length of its path to the root. We map the latent space of the model to the root tier (distance 0), and each output of the linear layer in the decoder aligns with a distinct hierarchical distance within the ontology. As a result, more general biological concepts are captured near the latent space, and specificity increases as we progress through the decoder toward the output layer, which consists of genes. Since DAGs often have a single root, a one-dimensional latent space is not informative enough. To improve model stability and interpretability, we introduce a preprocessing step that removes GO terms that are either overly broad (associated with many genes) or overly specific (linked to very few genes). GO terms associated with very few genes tend to yield unstable activity estimates in single-cell data due to sparsity and dropout effects, increasing the risk of overfitting through highly localized decoder connections. Conversely, terms associated with a large number of genes correspond to broad, high-level biological categories with limited discriminatory power across conditions. These terms introduce dense connections that increase computational cost while reducing interpretability. We adopt thresholds of <30 and >1000 genes, consistent with prior ontology-based models such as OntoVAE, where these values were empirically chosen to balance statistical robustness and biological interpretability. Threshold selection directly influences model behavior. Lowering the minimum threshold (e.g., <10 genes) increases sensitivity to noise and destabilizes GO activity estimation, whereas raising it excessively may eliminate biologically specific processes. Similarly, increasing the upper threshold (e.g., >2000 genes) introduces overly general terms that dominate the model and compress variation in GO activity, while setting it too low may remove important mid-level biological functions. The selected thresholds (<30 and >1000) strike a practical balance, preserving meaningful hierarchical structure while ensuring stable and interpretable performance. This pruning step also reduces model complexity and can alter the effective position of GO terms within the hierarchy, potentially changing their assignment to decoder layers.

In IST, the latent space has dimensions (n_1), which correspond to the number of GO terms at the root level of the trimmed GO tree. The first decoder layer contains one pytorch transformer layer, which takes the input of the latent space and produces an output with its dimension equal to that of the root level of GO trees. The IST decoder is composed of multiple transformer layers followed by linear layers. The linear layers reflect the hierarchical structure of the GO, progressing from more general biological processes to specific gene-level outputs. Each layer contains multiple transformers with transformed output of same dimension as input. The attention mechanism in the decoder is particularly important when integrating GO information because it helps the model selectively focus on biologically relevant features and makes sense of the complex hierarchical structure of gene function. The number of maximum linear layers depends on the hierarchical distance of the trimmed GO tree. The linear layer of the first decoder layer takes n_1 transformed output and produces n_2 outputs (Fig. 1A). Here, n_2 is the number of nodes that denote one-unit hierarchical distance in the trimmed ontology. The second decoder layer includes the concatenated nodes of the input and output of the first linear layer. That is, the total number of the input in the second layer becomes $n_1 + n_2 = n_3$. The second layer similarly goes through a transformer layer and linear layer and produces output n_4, which is the number of nodes with two-unit hierarchical distance to the root within the ontology. IST constructs all layers similarly. In this way, the input dimension of the transformer of the last decoder layer is n_{m-1}, which is the concatenation of previous linear layer's input dimension and output dimension with m represents a parameter related to the number of layers. The output dimension of the transformer is same size as input (i.e. n_{m-1}). This transformer output goes to the linear layer and produces n_m output because the last hierarchical distance within the ontology has n_m GO term in it. Then, the input and the output of the last linear layer of the decoder layer is concatenated ($n_{m-1}+n_m$) and are connected to the corresponding genes (output layer). In IST, node concatenation is used to enable the modeling of connections not only between adjacent layers but also between non-adjacent layers and the final output layer. This is accomplished through a progressive concatenation strategy, inspired by skip connections in DenseNet[32], where outputs from each previous decoder layer are appended to the current layer during the forward pass (Fig. 1B). Additionally, a binary mask is applied at each step to zero out weights corresponding to nonexistent connections, ensuring that only valid ontology-based relationships are retained. This approach allows us to represent potential connections between any pair of layers within the ontology (Fig. 1A, B).

Loss functions

Another novelty of IST is its loss function. This loss function aims to enable an accurate reconstruction of the original gene expression and a reliable interpretation of the biological processes. IST applies the Mean Squared Error (MSE) loss for accurate reconstruction of input gene expression profiles. To calculate this loss, the input gene expression vector $x = [x_1, x_2 \ldots \ldots x_n]$ is normalized as in (1).

$$x_i = \frac{x_i - min\,(x)}{max\,(x) - min\,(x)} \tag{1}$$

The MSE loss is then computed using (2), where y denotes the output gene expression from the model.

$$MSE\,(x, y) = \frac{1}{n}\sum\nolimits_{i=1}^{n}(x_i - y_i)^2 \tag{2}$$

To ensure biological relevance, we introduce a second loss, a correlation loss of the model-predicted GO rankings with the GO ranking obtained from enrichment tools such as EnrichR on training data[33]. We used Seurat in R which identifies differential expressed (DE) genes for different conditions of the three datasets. For example, in mouse skeletal muscle dataset DE genes were obtained between expression data at one condition (say day 0) and the data at other conditions (say day 2, 5, 7). Similarly, DE genes were obtained between stimulated and control

cells in the PBMC dataset and between stimulation time (0, 1, and 6 h) in the mouse MEFs dataset. These DE genes are then analyzed with EnrichR to perform GO enrichment analysis, producing a ranked list of GO terms based on corrected p-values (from smallest to largest). In parallel, IST infers the activities of every GO term in the trimmed ontology tree and then ranks the GO terms based on their activities from the largest to the smallest. The correlation loss is then calculated as one minus the Spearman's correlation of the ranks of the shared GOs. Thus, perfectly correlated predictions yield zero loss, while deviations from the EnrichR ranking increase the penalty.

IST adds the two loss functions together when calculating the overall loss. With this loss function, IST utilizes the Stochastic Gradient Descent optimizer[34], a standard and robust optimization algorithm, with a learning rate of 10^{-3}, to train the models to understand the biological process activities behind the expression data.

Comparisons with alternative models

To ensure a fair and transparent comparison, we designed our experiments by starting from existing ontology-based models and progressively introducing modifications. Our objective is not only to achieve accurate reconstruction of gene expression, but also to ensure that the model captures biologically meaningful gene activities reflect known biology across conditions.

We first implemented OntoVAE[12] as the baseline, which employs a GO-constrained architecture with a variational objective consisting of reconstruction loss and Kullback–Leibler (KL) divergence. While OntoVAE provides a principled probabilistic framework, it demonstrated limited ability to recover biologically meaningful GO activity patterns in our setting. We then extended this model to GO_VAE + correlation, incorporating a Spearman's correlation loss into the variational objective. This additional term aligns model-inferred GO activity rankings with those derived from EnrichR, encouraging prioritization of biologically relevant processes. However, this combination led to degraded performance, likely due to tension between the KL divergence, which enforces a smooth latent distribution, and the correlation loss, which preserves condition-specific biological variation. To address this issue, we removed the KL divergence term and developed GO_AE + correlation, an autoencoder trained with reconstruction loss MSE and correlation loss. This modification improved biological alignment, suggesting that relaxing the variational constraint enables better capture of meaningful biological variation.

Next, we evaluated two attention-based ablation models. In GO_AE + attention (decoder), transformer-based self-attention was introduced into the ontology-guided decoder while retaining only the reconstruction objective (MSE), isolating the contribution of attention within the GO hierarchy. In GO_AE + attention (encoder) + correlation, attention was instead applied in the encoder while maintaining the AE objective (MSE + correlation), allowing us to assess whether attention over gene features alone is sufficient to capture biological structure.

Finally, the full IST model, GO_AE + attention (decoder) + correlation, integrates GO-constrained architecture, decoder-level attention, and correlation-based supervision. This formulation enables dynamic modeling of interactions among biological processes while maintaining alignment with known biological signals.

Overall, this progressive framework allows us to systematically disentangle the contributions of attention mechanisms and correlation-based supervision, ensuring controlled and interpretable comparisons across models.

Results

IST reliably measures condition-specific biological activities

To evaluate the ability of IST to infer condition-specific biological activities from scRNA-seq data, we compared the rankings of GO terms inferred by IST with the rankings of enriched GO terms derived from EnrichR analysis of DE genes in the testing data (Material and methods, Table 1). IST infers the biological activities through the weights of the GO terms in its attention-based decoder and ranks GO terms based on their weights from the largest to the smallest. For EnrichR GO ranks, we used Seurat to obtain DE genes between a stage and all other stages in the mouse skeletal muscle dataset, between one time to all other times in mouse MEF dataset, and between control and treatment in the human dataset. We then used EnrichR to obtain and rank the biological process GOs of a stage or a condition from the most to the least enriched GO terms.

Dataset		No. of GO terms			Spearman's Correlation (*r*)	
		EnrichR	IST	Common	EnrichR & IST	EnrichR & Shuffle IST
Injury stages	Day 0	5303	2978	2061	0.949	0.927
	Day 2	5406	2978	2062	0.947	0.929
	Day 5	5308	2978	2061	0.950	0.929
	Day 7	5286	2978	2062	0.922	0.905
PBMC data		5081	3042	2278	0.945	0.908
Mouse MEFs	0 h	3809	3300	1802	0.861	0.801
	1 h	5249	3300	2398	0.951	0.865
	6 h	5269	3300	2405	0.890	0.828

Table 1. Correlation analysis of IST's GO ranking.

We found that Spearman's correlation of the two ranks on the mouse skeletal muscle dataset at each of the four stages ranged from 0.922 to 0.95, when inputting expression data of any stage without telling the stage origin of the input data (Table 1). We also obtained similar Spearman's correlation (0.945) for human testing data and mouse MEFs (0.86 to 0.95). Note that although IST used EnrichR GO rank of the training data during training, it does not use EnrichR GO rank of the testing data. The above high correlation thus indicated that IST reliably captures true biological activities that characterize important genes at different stages.

To assess the model's robustness, we examined how perturbations in the EnrichR GO ranking affected IST's performance. Specifically, we retrained the same IST architecture with the shuffled EnrichR GO rank, where 5% of GO terms were randomly switched positions in the EnrichR GO ranking. In this way, we generated ten Shuffle IST models. All models were evaluated on identical testing data. Comparative analyses revealed that correlations of GO rank between Shuffle IST and EnrichR decreased only slightly on average. For instance, at stage 0 of the mouse skeletal muscle, the average correlation of the GO rank of the Shuffle IST model and that of EnrichR was 0.923, while the correlation was 0.949 for the original IST model and EnrichR (Table 1). Such high correlations across the three datasets suggest that the IST model is insensitive to the input EnrichR GO rank.

Collectively, these results support that IST maintains a robust internal representation of true biological activity. Rather than passively inheriting rankings from EnrichR, the model learns biologically consistent organization across stages and treatments.

IST identifies more context-specific biological activities than EnrichR

We showed that the inferred context-specific GO activities captured the activities of DE genes. In practice, the inferred context-specific activities should also represent the activities of non-DE genes. Indeed, IST infers context-specific GO activities of both DE and non-DE genes that are important under a specific experimental condition.

We already showed the power of IST's identification of representative activities of DE genes at each stage above. Here, we checked whether the top enriched GOs of the DE genes were included in top GO activities reported by IST under a condition. We found that this was indeed true. For instance, the top three EnrichR GO terms across the four mouse skeletal muscle injury stages were included in top 30 GO terms of IST, suggesting that the obtained IST GO terms are biologically significant.

We further checked where the top three IST GO terms were in the list of EnrichR GO terms. EnrichR predicted 392, 390, 344, and 298 GO terms with corrected p-value < 0.01 in days 0, 2, 5 and 7, respectively, in mouse skeletal muscle dataset by the Benjamini-Hochberg procedure[35]. Although the IST top three GO terms at all stages appeared within EnrichR's top 100 GO terms, there were 0 to 21.7% non-DE genes that contributed to the activities of a specific biological process GO term. The details were in the following:

Day 0 (Baseline or Uninjured State): IST ranked protein localization (GO:0008104), regulation of cellular component organization (GO:0051128), and small GTPase-mediated signal transduction (GO:0007264) as the top three activities at Stage 0, while EnrichR ranked them at 28, 90, and 25, respectively. There were 11.62%, 0%, and 4.79% non-DE genes that contributed to these top GO activities, respectively. These top activities makes perfect sense, because proper spatial distribution of proteins and organelles is essential for preserving sarcomere integrity and supporting the quiescent state of satellite cells[36]. Likewise, Rho/Rac/Cdc42 GTPase pathways regulate cytoskeletal architecture, adhesion, and niche interactions, enabling stem cells to remain poised for rapid activation upon injury[37].

Day 2 (Early Activation and Proliferation): At this stage, IST had protein localization (GO:0008104), small GTPase-mediated signal transduction (GO:0007264), and regulation of cell population proliferation (GO:0042127) as the top three biological activities, while EnrichR ranked them at 22, 25, and 61, respectively. There were 5.56%, 5.49%, and 7.06% non-DE genes contributing to the three activities, respectively. The prominence of these top three activities matches the well-established burst of myogenic precursor amplification that peaks during this early repair phase, because satellite cells have exited quiescence and entered a proliferative program, which is captured by the enrichment of regulation of cell population proliferation[38] alongside continued signals for protein localization[36] and small GTPase-mediated signaling[37]. Moreover, early activation requires major rearrangements in cell polarity, receptor trafficking, and cytoskeletal remodeling, activities tightly regulated by GTPase pathways.

Day 5 (Growth Regulation and Differentiation): At this phase, IST reported protein localization (GO:0008104), positive regulation of cellular component biogenesis (GO:0044089), and regulation of angiogenesis (GO:0045765) as the top three activities, while EnrichR ranked them at 39, 139, and 42, respectively. There were 12.63%, 13.51%, and 8.33% non-DE genes that contributed to the three activities, respectively. These inferred top activities capture the transition toward differentiation through enrichment of cellular component biogenesis and regulation of angiogenesis. During this stage, proliferating myoblasts fuse to form nascent myofibers, driving substantial increases in biosynthetic demand, organelle assembly, and structural protein production[39]. Concurrently, angiogenic programs become highly active, supporting the metabolic needs of regenerating fibers and coordinating cross-talk between endothelial and myogenic cells[40]. The recurrence of protein localization further reflects active trafficking and assembly processes required for sarcomere formation and maturation.

Day 7 (Remodeling and Maturation): IST ranked positive regulation of autophagy (GO:0010506), small GTPase-mediated signal transduction (GO:0007264), and regulation of cell cycle (GO:0051726) as the top three terms, while EnrichR gave them ranks 7, 9, 30, respectively. There were 13.21%, 15.57%, and 17.70% non-DE genes that contributed to the three GO activities, respectively. Autophagy becomes critical for removing damaged organelles, reshaping metabolic programs, and enabling proper differentiation and long-term tissue maintenance[41]. Continued enrichment of small GTPase pathways indicates ongoing cytoskeletal reorganization necessary for sarcomere alignment and membrane stabilization in maturing myofibers. Meanwhile, control of the cell cycle reflects the coordinated exit of myoblasts from proliferation and the reestablishment of the satellite

cell pool through self-renewal. The inferred top three activities at this phase thus define the late-stage transition toward functional muscle restoration.

Similarly, the top three activities inferred by IST on human PBMC and mouse MEFs datasets were also supported by literature. The results on three different datasets demonstrate that IST not only recapitulates but refines biological inference from scRNA-seq data, offering mechanistically consistent interpretations of biological activities.

IST identifies dependency of GO terms

To characterize how GO terms are functionally related, we examined the cross-GO attention patterns generated by the model. Because the full attention matrix contains a large number of GO-GO interactions, we focused on the top 100 GO pairs with the highest attention weights. In the mouse skeletal muscle dataset, none of these highly weighted pairs displayed a direct parent-child or sibling relationship. When we expanded the analysis to include GO terms sharing a common ancestor within four hierarchical levels, we found that approximately 50% of pairs at day 0, 47% at days 2 and 5, and 37% at day 7 shared such ancestry. A similar trend was observed in other two datasets: none of the top 100 pairs exhibited direct hierarchical links, and only 37% shared a common ancestor when we allowed a broader distance of four to eight levels. Together, these results indicate that IST assigns high attention not to terms with strict GO-structural proximity, but to functionally related processes, suggesting that the model captures higher-order biological organization rather than local GO hierarchy.

We further focused on the top attention weights of top-ranked GO terms based on IST's GO activity rank. In the mouse skeletal muscle dataset, the top GO term at day 0 was protein localization (GO:0008104). Two GO terms: manganese ion transmembrane transport (GO:0071421) and negative regulation of protein localization to cilium (GO:1903565), received the highest attention weights toward it. Although GO:0071421 is not hierarchically related to GO:0008104, manganese transport is an upstream biochemical requirement for components of the protein-localization machinery[42,43]. In contrast, GO:1,903,565 is both hierarchically and mechanistically linked to GO:0008104[44]. Similarly, in day 2 to day 7, both the above GO terms (GO:0071421 and GO:1903565) showed consistently high attention weights for top GOs that were functionally related based on literature[42,44,45]. In the human dataset, IST's top GO term was modification-dependent protein catabolic process (GO:0019941), which includes ubiquitin-mediated proteasomal degradation. Many cellular pathways either rely on or regulate this degradation system, forming functional connections that the model captures through attention[46]. One prominent attention hit was regulation of metal ion transport (GO:0010959). Although not a direct parent/child in GO structure, metal ion homeostasis strongly influences proteostasis and stress-response pathways that activate ubiquitin-dependent degradation[47]. Similar trend was observed in mouse MEF dataset. Thus, the transformer correctly identifies functional coupling even in the absence of direct GO-hierarchy links.

IST predicts new functions of genes

Beyond inferring biological activity from transcriptomic data, IST can predict new functions of genes. To explore this capability, we performed computational gene knockout experiments. Specifically, for each gene under a condition, we set its expression to zero and re-estimated the corresponding biological activity profile. We then compared these profiles with those obtained from the unmodified data. Genes whose knockouts produced significant shifts in activity (Wilcoxon test, false discovery rate < 0.01) were designated as important. These important genes showed strong overlap with DE genes across datasets (for instance, mouse regeneration stages Day 0: 92.86%, Day 2: 91.93%, Day 5: 87.04%, Day 7: 82.21%; PBMCs: 68.74%), supporting the biological reliability of IST predictions.

We next examined how GO term activities change in response to individual gene knockouts across conditions. For each knockout, we identified significantly altered GO activities using a Wilcoxon test with multiple-testing correction (adjusted $p\text{-}value < 0.01$). GO terms with significant changes were interpreted as the putative functional roles of the perturbed gene.

To assess the agreement between IST-inferred functions and existing knowledge, we compared the predicted GO terms with each gene's annotated GO terms within the trimmed ontology and computed the percentage of overlap. IST achieved approximately 60% overlap in the mouse skeletal muscle and mouse MEF datasets, and 40% in the PBMC dataset, indicating strong concordance with established gene functions. To evaluate whether this agreement exceeds random expectation, we performed a control experiment in which GO terms were randomly reassigned to genes while preserving the original gene expression data, and the models were retrained. Under this setting, the overlap dropped to ~16% for the mouse skeletal muscle and MEF datasets, and ~25% for the PBMC dataset. These substantially lower overlaps suggest that IST captures biologically meaningful gene-function relationships and that its predictions of novel functions are unlikely to arise by chance. The nonzero overlap in the randomized setting may be attributable to shared or broadly annotated GO terms, as well as dataset-specific factors such as the smaller sample size of the PBMC data.

Importantly, IST also revealed novel functions for a given gene. To assess the biological relevance of these newly inferred functions, we examined a representative example, Pecam1 (CD31), across all mouse skeletal muscle injury stages. Pecam1 was selected because it was identified as an important gene by both IST and Seurat, and its role in endothelial biology is well established[29]. The in silico knockout of Pecam1 revealed new GO terms consistent with its known biological roles like vascular integrity[29] and endothelial junction formation[48]. For instance, the unannotated function, regulation of acute-phase response (GO:0006953) was significantly altered even at the uninjured Day 0, reflecting endothelial activation. At Day 5, several vascular-related terms, like negative regulation of vascular permeability (GO:0043116)[49], positive regulation of vascular permeability (GO:0043117)[50], and amoeboid-type cell migration (GO:0001667)[51], showed highly significant changes ($p \approx 0$). These terms correspond to known Pecam1-mediated processes involved in maintaining endothelial barrier integrity, regulating permeability during inflammation, and facilitating leukocyte and endothelial cell migration

during angiogenesis[52]. Together, these findings support our hypothesis that IST can be used to predict new functions of genes.

IST shows superior performance to other models

To benchmark the performance of IST, we compared it against multiple model variants (Material and Methods). We evaluated each model using both predictive performance accuracy along with expression reconstruction error which is the mean absolute error (MAE) and biological label correctness (stage, time point, or condition of the sample in the corresponding dataset) (Table 2).

IST consistently achieves better overall performance across all datasets, demonstrating high predictive accuracy, strong alignment with biological labels, and low reconstruction error (Table 2). In contrast, although OntoVAE attains the lowest reconstruction error, this does not translate into reliable biological labeling performance, indicating that optimizing reconstruction alone is insufficient to yield biologically meaningful or discriminative representations. The comparison between GO_VAE + correlation and GO_AE + correlation further highlights the impact of the KL divergence term. While incorporating correlation loss into the VAE objective provides limited benefit, removing the KL divergence (GO_AE + correlation) leads to substantial improvements in both predictive accuracy and biological alignment. This suggests that the variational constraint may hinder the preservation of condition-specific biological signals.

The ablation studies further clarify the role of attention. The GO_AE + attention (decoder) model yields only moderate gains, indicating that attention alone is insufficient without biologically informed supervision. Similarly, GO_AE + attention (encoder) + correlation achieves competitive performance in some datasets but remains inconsistent overall, suggesting that attention applied solely at the gene level does not fully capture structured biological relationships. In contrast, IST incorporates attention within the ontology-guided decoder, enabling dynamic modeling of interactions among GO terms while being guided by correlation-based supervision. This integration results in the most robust and biologically consistent performance across all datasets. Collectively, these findings demonstrate that ontology structure, correlation-based supervision, and decoder-level attention act synergistically, and that their integration in IST is critical for achieving superior and biologically meaningful performance.

We further evaluated clustering performance using GO activity representations derived from each model. Applying Leiden clustering (resolution = 0.25) to the mouse skeletal muscle dataset, where cell-type labels are available, IST achieved an adjusted Rand index (ARI) of 0.55, outperforming all competing models (Table 3). Notably, although IST and GO_AE + attention (encoder) + correlation exhibit comparable correlations with EnrichR (Table 2), their downstream clustering performance differs substantially. This indicates that IST not only ranks GO terms more meaningfully but also produces representations that better separate biologically distinct cell populations during regeneration.

Time complexity analysis

From a theoretical standpoint, IST scales linearly with the number of cells and genes, but quadratically with the number of retained GO terms due to the self-attention mechanism in the decoder. The encoder consists of linear layers and has a computational complexity of approximately $O(C \cdot G)$, where C is the number of cells and G the number of genes. This is comparable to standard AEs and VAEs. As training is performed using mini-batch

Dataset	Model variant	GO structure	Attention	Correlation loss	Average correlation score	Expression reconstruction MAE	Prediction accuracy
Mouse skeletal muscle	OntoVAE	Yes	No	No	0.001	0.003	85.069%
	GO_VAE+ correlation	Yes	No	Yes	0.054	0.012	77.091%
	GO_AE+ correlation	Yes	No	Yes	0.870	0.011	97.400%
	GO_AE+ attention	Yes	Yes (Decoder)	No	0.200	0.012	88.439%
	GO_AE+ attention	Yes	Yes (Encoder)	Yes	0.942	0.011	94.171%
	IST (GO_AE+ attention in Decoder)	**Yes**	**Yes (Decoder)**	**Yes**	**0.942**	**0.011**	**97.300%**
Human PBMC	OntoVAE	Yes	No	No	-0.015	0.001	50.050%
	GO_VAE+ correlation	Yes	No	Yes	0.095	0.018	84.167%
	GO_AE+ correlation	Yes	No	Yes	0.945	0.018	86.002%
	GO_AE+ attention	Yes	Yes (Decoder)	No	0.012	0.018	78.101%
	GO_AE+ attention	Yes	Yes (Encoder)	Yes	0.945	0.015	86.967%
	IST (GO_AE+ attention in Decoder)	**Yes**	**Yes (Decoder)**	**Yes**	**0.945**	**0.016**	**91.489%**
Mouse MEFs	OntoVAE	Yes	No	No	0.016	0.001	98.693%
	GO_VAE+ correlation	Yes	No	Yes	0.901	0.015	93.698%
	GO_AE+ correlation	Yes	No	Yes	0.901	0.014	98.312%
	GO_AE+ attention	Yes	Yes (Decoder)	No	0.019	0.015	52.595%
	GO_AE+ attention	Yes	Yes (Encoder)	Yes	0.900	0.015	56.943%
	IST (GO_AE+ attention in Decoder)	**Yes**	**Yes (Decoder)**	**Yes**	**0.901**	**0.016**	**99.071%**

Table 2. Model comparison on three datasets.

	IST		GO_VAE+ correlation		GO_AE+ correlation		GO_AE+ attention (decoder)		GO_ AE+attention (encoder) + correlation		OntoVAE	
Injury stage	*r*	ARI	*r*	ARI	*r*	ARI	*r*	ARI	*r*	ARI	*r*	ARI
Day 0	0.949	0.55	0.053	0.24	0.942	0.47	0.193	0.24	0.949	0.48	0.001	0.48
Day 2	0.947		0.049		0.912		0.192		0.949		− 0.001	
Day 5	0.950		0.063		0.936		0.238		0.948		0.009	
Day 7	0.922		0.049		0.690		0.177		0.922		− 0.023	

Table 3. Spearman's correlations (*r*) with EnrichR GO ranks and ARI index of different models.

Model	Complexity	Training time (per epoch)
GO_VAE+ correlation	$O(C.G)$	~7 s
GO_AE+ correlation	$O(C.G)$	~7 s
GO_AE+ attention (decoder)	$O(C.G)$	~6 s
GO_AE+ attention (encoder)+correlation	$O(C.G) + O(G^2)$	~11 s
OntoVAE (GO_VAE)	$O(C.G)$	~4 s
IST (proposed)	$O(C.G) + O(T^2)$	~12 s

Table 4. Time complexity of the models.

stochastic gradient descent, increasing the number of cells impacts total training time linearly while having minimal effect on per-batch memory usage.

The primary computational overhead arises from the decoder's transformer-based attention over GO nodes, which scales as $O\left(T^2\right)$, where T is the number of retained GO terms. In practice, after ontology filtering, we retain approximately 3,000 GO terms (e.g., 2,978 in mouse injury, 3,042 in PBMC, and 3,300 in mouse MEFs), keeping the attention matrix tractable on modern GPUs. Notably, attention is applied only across GO nodes, not across cells. As a result, IST avoids the quadratic scaling with respect to cell number observed in cell-attention models, enabling efficient application to large-scale datasets via mini-batch training. Empirically, IST incurs moderately higher training time than simpler baselines due to the inclusion of attention layers and ontology masking, but remains computationally practical. To quantify this, we compared training time and memory usage with baseline models under a consistent hardware setting (NVIDIA A100 80GB GPU, 1.97 TB RAM) using the human PBMC dataset (Table 4).

Overall, IST introduces modest computational overhead relative to standard AEs and VAEs, yet remains comparable to other transformer-based architectures. Crucially, because IST does not perform attention across cells, it can scale to atlas-level datasets comprising hundreds of thousands to millions of cells without requiring architectural modifications.

Discussion and conclusion

IST establishes a framework for embedding biological priors into attention-based architectures, bridging predictive modeling with mechanistic interpretability. IST provides a biologically grounded and generalizable approach for scRNA-seq analysis.

The results obtained from IST highlight its ability to capture biologically meaningful GO terms from high-dimensional scRNA-seq data in both human and mouse data. Through the integration of GO knowledge and attention mechanisms, the model not only enhances predictive performance but also provides interpretable insights into gene functions. This interpretability is particularly valuable for identifying key pathways, regulatory programs, or potential therapeutic targets. A key strength of the GO-guided attention design is its alignment with biological prior knowledge, which grounds the learned representations in established functional hierarchies. By weighing genes and GOs according to their biological relevance, IST offers an interpretable framework that moves beyond purely data-driven latent embeddings. This enables to better connect machine learning outputs with established biological processes, thereby fostering more trust and adoption of such models in biomedical research. Unlike traditional black-box deep learning models, IST promotes transparency, supporting hypothesis generation and validation in experimental settings. In contrast to topology-driven approaches such as scMGCA[53], where interpretability emerges indirectly from preserved cell-cell relationships and is typically inferred post hoc through marker identification or enrichment analysis, IST adopts a top-down strategy by explicitly embedding biological priors into the model architecture. This design enables IST to generate GO activities as explicit intermediate representations during training, providing more direct, structured, and mechanistically grounded interpretability compared to purely data-driven latent embeddings.

Nevertheless, some limitations still remain. Firstly, the reliance on the accuracy and completeness of GO annotations introduces potential biases, as incomplete or outdated ontologies may constrain model interpretability. Secondly, the performance of IST is sensitive to dataset size and quality; small or noisy datasets may limit the model's generalization and stability. While we evaluated IST on three datasets to demonstrate its

broader applicability, further validation across a much wider range of datasets spanning different tissues, species, technologies, and batch conditions would be necessary to fully establish the generalizability of the framework. Finally, the training is dependent on the GO enrichment analysis ranking for guiding the model to obtain better prediction. Future work could explore models without this dependency. Expanding the model toward clinical applications, such as biomarker discovery or patient stratification, also presents an exciting direction for translational research.

In summary, IST demonstrates how embedding biological priors into attention-based VAE can yield models that are both powerful and interpretable, offering a meaningful step forward in merging machine learning with systems biology for transcriptomic data analysis.

Data availability

The datasets and tool are available at https://doi.org/10.6084/m9.figshare.30811334 and https://github.com/TasbirahaAthaya/IST.

Received: 1 February 2026; Accepted: 17 April 2026
Published online: 24 April 2026

References

1. Ge, S., Sun, S., Xu, H., Cheng, Q. & Ren, Z. Deep learning in single-cell and spatial transcriptomics data analysis: advances and challenges from a data science perspective. *Brief. Bioinform*. **26**, bbaf136 (2025).
2. Molho, D. et al. Deep Learning in Single-cell Analysis. *ACM Trans. Intell. Syst. Technol.* **15** (1–40), 62 (2024).
3. Athaya, T., Ripan, R. C., Li, X. & Hu, H. Multimodal deep learning approaches for single-cell multi-omics data integration. *Brief. Bioinform.* **24**, bbad313 (2023).
4. Luo, Z., Xu, C., Zhang, Z. & Jin, W. A topology-preserving dimensionality reduction method for single-cell RNA-seq data using graph autoencoder. *Sci. Rep.* **11**, 20028 (2021).
5. Athaya, T., Li, X. & Hu, H. A deep learning method to integrate extracelluar miRNA with mRNA for cancer studies. *Bioinformatics* **40**, btae653 (2024).
6. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* **10**, 390 (2019).
7. Tran, D. et al. Fast and precise single-cell data analysis using a hierarchical autoencoder. *Nat. Commun.* **12**, 1029 (2021).
8. Shi, X., Ramathal, C. & Dezso, Z. I. N. S. I. G. H. T. In Silico Drug Screening Platform using Interpretable Deep Learning Network. 04.21.649855 Preprint at (2025). https://doi.org/10.1101/2025.04.21.649855 (2025).
9. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nat. Methods*. **16**, 715–721 (2019).
10. Dong, J., Zhang, Y. & Wang, F. scSemiAE: a deep model with semi-supervised learning for single-cell transcriptomics. *BMC Bioinform.* **23**, 161 (2022).
11. Svensson, V., Gayoso, A., Yosef, N. & Pachter, L. Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics* **36**, 3418–3421 (2020).
12. Doncevic, D. & Herrmann, C. Biologically informed variational autoencoders allow predictive modeling of genetic and drug-induced perturbations. *Bioinformatics* **39**, btad387 (2023).
13. Ribeiro, M. T., Singh, S., Guestrin, C. & 'Why Should, I. Trust You?': Explaining the Predictions of Any Classifier. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135–1144Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2939672.2939778 (2016).
14. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. in *Proceedings of the 31st International Conference on Neural Information Processing Systems* 4768–4777Curran Associates Inc., Red Hook, NY, USA, (2017).
15. Lotfollahi, M. et al. Biologically informed deep learning to query gene programs in single-cell atlases. *Nat. Cell. Biol.* **25**, 337–350 (2023).
16. Seninge, L., Anastopoulos, I., Ding, H. & Stuart, J. VEGA is an interpretable generative model for inferring biological network activity in single-cell transcriptomics. *Nat. Commun.* **12**, 5684 (2021).
17. Ma, J. et al. Using deep learning to model the hierarchical structure and function of a cell. *Nat. Methods*. **15**, 290–298 (2018).
18. Elmarakeby, H. A. et al. Biologically informed deep neural network for prostate cancer discovery. *Nature* **598**, 348–352 (2021).
19. Kuenzi, B. M. et al. Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells. *Cancer Cell.* **38**, 672–684e6 (2020).
20. Doncevic, D. et al. Prediction of context-specific regulatory programs and pathways using interpretable deep learning. 11.06.622202 Preprint at (2024). https://doi.org/10.1101/2024.11.06.622202 (2024).
21. Gillioz, A., Casas, J., Mugellini, E. & Khaled, O. A. Overview of the Transformer-based Models for NLP Tasks. In *15th Conference on Computer Science and Information Systems (FedCSIS)* 179–183 (2020). https://doi.org/10.15439/2020F20.
22. Qi, Y., Kan, Y., Qi, J. & Jin, S. scGT:integration algorithm for single-cell RNA-seq and ATAC-seq based on graph transformer. *Bioinformatics* **41**, btaf357 (2025).
23. Liu, X. et al. Pathformer: a biological pathway informed transformer for disease diagnosis and prognosis using multi-omics data. *Bioinformatics* **40**, btae316 (2024).
24. Nouri, N. Single-cell RNA-seq data augmentation using generative Fourier transformer. *Commun. Biol.* **8**, 113 (2025).
25. Qin, H., Shi, X., Zhou, H. & scSwinFormer: A Transformer-Based Cell-Type Annotation Method for scRNA-Seq Data Using Smooth Gene Embedding and Global Features. *J. Chem. Inf. Model.* **64**, 6316–6323 (2024).
26. Chen, J. et al. Transformer for one stop interpretable cell type annotation. *Nat. Commun.* **14**, 223 (2023).
27. Szałata, A. et al. Transformers in single-cell omics: a review and new perspectives. *Nat. Methods*. **21**, 1430–1443 (2024).
28. Zheng, H., Wang, S., Li, X. & Hu, H. INSISTC: Incorporating network structure information for single-cell type classification. *Genomics* **114**, 110480 (2022).
29. De Micheli, A. J. et al. Single-Cell Analysis of the Muscle Stem Cell Hierarchy Identifies Heterotypic Communication Signals Involved in Skeletal Muscle Regeneration. *Cell. Rep.* **30**, 3583–3595e5 (2020).
30. Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
31. Wang, X. & Zhou, L. The Many Roles of Macrophages in Skeletal Muscle Injury and Repair. *Front Cell. Dev. Biol* **10**, (2022).
32. Huang, G., Liu, Z., Maaten, L. V. D. & Weinberger, K. Q. Densely Connected Convolutional Networks. *IEEE Comput. Soc.* 2261–2269. https://doi.org/10.1109/CVPR.2017.243 (2017).
33. Chen, E. Y. et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinform.* **14**, 128 (2013).
34. Kiefer, J. & Wolfowitz, J. Stochastic Estimation of the Maximum of a Regression Function. *Ann. Math. Stat.* **23**, 462–466 (1952).
35. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Royal Stat. Soc. J. Ser. B: Methodological*. **57**, 289–300 (1995).

36. Anderson, J. E. Key concepts in muscle regeneration: muscle cellular ecology integrates a gestalt of cellular cross-talk, motility, and activity to remodel structure and restore function. *Eur. J. Appl. Physiol.* **122**, 273–300 (2022).
37. Kann, A. P. et al. An injury-responsive Rac-to-Rho GTPase switch drives activation of muscle stem cells through rapid cytoskeletal remodeling. *Cell. Stem Cell.* **29**, 933–947e6 (2022).
38. Careccia, G., Mangiavini, L. & Cirillo, F. Regulation of Satellite Cells Functions during Skeletal Muscle Regeneration: A Critical Step in Physiological and Pathological Conditions. *Int. J. Mol. Sci.* **25**, 512 (2024).
39. Hees, J. T. & Harbauer, A. B. Metabolic Regulation of Mitochondrial Protein Biogenesis from a Neuronal Perspective. *Biomolecules* **12**, 1595 (2022).
40. Koike, H., Manabe, I. & Oishi, Y. Mechanisms of cooperative cell-cell interactions in skeletal muscle regeneration. *Inflamm. Regener.* **42**, 48 (2022).
41. Chu, Y., Yuan, X., Tao, Y., Yang, B. & Luo, J. Autophagy in Muscle Regeneration: Mechanisms, Targets, and Therapeutic Perspective. *Int. J. Mol. Sci.* **25**, 11901 (2024).
42. Legrand, D. et al. New insights into the pathogenicity of TMEM165 variants using structural modeling based on AlphaFold 2 predictions. *Comput. Struct. Biotechnol. J.* **21**, 3424–3436 (2023).
43. Dulary, E., Potelle, S., Legrand, D. & Foulquier, F. TMEM165 deficiencies in Congenital Disorders of Glycosylation type II (CDG-II): Clues and evidences for roles of the protein in Golgi functions and ion homeostasis. *Tissue Cell.* **49**, 150–156 (2017).
44. Palla, A. R. et al. Primary cilia on muscle stem cells are critical to maintain regenerative capacity and are lost during aging. *Nat. Commun.* **13**, 1439 (2022).
45. Gordon, S. J. V., Fenker, D. E., Vest, K. E. & Padilla-Benavides, T. Manganese influx and expression of ZIP8 is essential in primary myoblasts and contributes to activation of SOD2†. *Metallomics* **11**, 1140–1153 (2019).
46. Liu, L. et al. UbiHub: a data hub for the explorers of ubiquitination pathways. *Bioinformatics* **35**, 2882–2884 (2019).
47. Shenoy, A., Kalakoti, Y., Sundar, D. & Elofsson, A. M-Ionic: prediction of metal-ion-binding sites from sequence using residue embeddings. *Bioinformatics* **40**, btad782 (2024).
48. Kiseleva, R. Y. et al. Vascular endothelial effects of collaborative binding to platelet/endothelial cell adhesion molecule-1 (PECAM-1). *Sci. Rep.* **8**, 1510 (2018).
49. Lertkiatmongkol, P., Liao, D., Mei, H., Hu, Y. & Newman, P. J. Endothelial functions of platelet/endothelial cell adhesion molecule-1 (CD31). *Curr. Opin. Hematol.* **23**, 253 (2016).
50. Graesser, D. et al. Altered vascular permeability and early onset of experimental autoimmune encephalomyelitis in PECAM-1–deficient mice. *J. Clin. Invest.* **109**, 383–392 (2002).
51. Privratsky, J. R., Newman, D. K. & Newman, P. J. PECAM-1: Conflicts of interest in inflammation. *Life Sci.* **87**, 69–82 (2010).
52. Tusavitz, S. et al. Macrophage-derived Wnt signaling increases endothelial permeability during skeletal muscle injury. *Inflamm. Res.* **69**, 1235–1244 (2020).
53. Yu, Z. et al. Topological identification and interpretation for single-cell gene regulation elucidation across multiple platforms using scMGCA. *Nat. Commun.* **14**, 400 (2023).

Acknowledgements

We thank Matthew Weston from the Hu Lab for assistance with manuscript editing and formatting.

Author contributions

H.H. and X.L. conceived and designed the study. T.A. performed the experiments. T. A., X.L., and H.H. analyzed the data. T. A., X.L., and H.H. wrote the manuscript. All authors proofread the manuscript.

Funding

This work has been supported by the U.S. National Science Foundation grants 2120907 and 2514869.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to X.L. or H.H.

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

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