TFvelo: gene regulation inspired RNA velocity estimation

Jiachen Li[1], Xiaoyong Pan[1], Ye Yuan[1]*, Hong-Bin Shen[1]*
[1]Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, 200240, China
*Address correspondence to Y.Y at: yuanye_auto@sjtu.edu.cn, H.B. Shen at: hbshen@sjtu.edu.cn

Abstract

RNA velocity is closely related with cell fate and is an important indicator for the prediction of cell states with elegant physical explanation derived from single-cell RNA-seq data. However, most existing RNA velocity models can only be applied to datasets with unspliced/spliced or new/total RNA abundance information. Moreover, the dynamics constructed by these models can not fit the co-expression of spliced and unspliced RNAs well. Motivated by the finding that RNA velocity could be driven by the transcriptional regulation, we propose TFvelo, which expands RNA velocity concept to various single-cell datasets without splicing information, by introducing gene regulatory network information. Our experiments on synthetic data and scRNA-Seq data demonstrate that TFvelo can better model the gene dynamics, infer cell pseudo-time and trajectory, and also detect the key TF-target regulation simultaneously. TFvelo opens a novel, robust and accurate avenue for modeling RNA velocity for single cell data.

Introduction

Single-cell RNA sequencing (scRNA-seq) [1-3] provides a wealth of information about the gene expression profile of individual cells. To expand the scope of scRNA-seq analysis beyond a static snapshot and capture cellular dynamics without tracking alive individual cells over time, RNA velocity [4] was proposed. RNA velocity describes the derivative of gene expression over time, which offers a physical-informed approach to studying gene dynamics by modeling the relationship between the abundance of unspliced (immature) and spliced (mature) RNAs. In addition, by combining velocities across multiple genes, it is possible to infer the fate of individual cells. To estimate the RNA velocity, velocyto [4] is introduced in the vanilla approach with a steady state model. ScVelo [5] models the dynamics without the steady state assumption and employs an Expectation-Maximization (EM) approach for better reconstructing the underlying kinetics.

More recently, several approaches have been developed to improve the estimation of RNA velocity and the accuracy of the following cell transition prediction. For instance,

VeloAE [6] utilized an auto-encoder and low-dimensional space to smooth RNA velocity. Still in the unspliced/spliced space, DeepCycle [7] used an autoencoder to map cell cycle-related genes. UniTVelo [8] designed a function of time directly to model the spliced RNA level, and one more dynamic equation to model the relationship between the unspliced and spliced. In contrast to time-invariant and deterministic models, Pyrovelocity [9] and veloVAE [10] estimated RNA velocity using Bayesian inference frameworks. Furthermore, apart from the unspliced/spliced data, dynamics information can also be provided by new omics technologies. For instance, Dynamo [11] construct a RNA velocity model based on the new/total labeled RNA-seq [12]. Taking the advantage of single-cell multi-omics technology [13, 14] , RNA velocity analysis can be expanded to protein abundances (protaccel [15]) and single-cell ATAC-seq (MultiVelo [16]) datasets. However, it is cost and time-consuming to obtain these new omics data with additional labeling.

While RNA velocity theory has improved the inference of single-cell trajectories, pseudo-time, and gene regulation inference in numerous studies [17], current RNA velocity models still face severe challenges. Firstly, RNA velocity models always treat each gene independently and do not take the underlying regulation into consideration [18], although it is expected to be helpful for cell fate inference by integrating gene regulatory mechanism analysis. Secondly, the RNA velocity models can only be applied to scRNA-seq datasets with unspliced/spliced or new/total labels. However, transcriptional dynamics of mRNA splicing may not provide sufficient signal, especially in single cell resolution [18]. Last but not least, these models can not fit the gene dynamics well on most genes. That is to say, these approaches assume that the joint plot between unspliced and spliced expression should form a clockwise curve distribution [4] [5] [19], which is, however, rarely observed in the results.

Our study aims to uncover the underlying regulation among genes to better infer cell fate. To this end, we investigate the relationship between gene regulatory patterns and RNA velocity, which reveals that the RNA velocity can be represented as a linear combination of the expression levels of transcription factors (TFs). Notably, the TFs with the positive weights are significantly enriched with TF set functionally linked to the target gene. In addition, it is found that the joint distribution of expression between a TF and its target could form the clockwise curve distribution [20], indicating the phase delay between them. Those findings imply that, instead of unspliced-spliced information, the expression level of TF and its target could also be used to construct dynamics model of RNA velocity.

Here we propose a gene regulation inspired RNA velocity estimation method, TFvelo to address the three challenges mentioned above. Firstly, TFvelo is a novel method that models the RNA velocity of a target gene based on the expression levels of its identified TFs. The computational framework of TFvelo relies on a generalized EM algorithm, which iteratively updates the weights of the TFs, the latent time of cells, and the parameters in the dynamic equation. Secondly, this approach enables the application of

RNA velocity theory to various single-cell data without splicing information, new/total labelling. Finally, using both synthetic data and real datasets, we demonstrate that the TFvelo model can effectively capture the clockwise phase portraits of TF-target, accurately infer the pseudo-time and the developmental trajectories of cells, which can further help find key underlying TF-target regulation mechanisms.

Results

Findings: RNA velocity can be modelled by TFs' expression level.

The fact that RNA velocity can be modelled by the transcriptional regulation is supported by the following two findings. Firstly, we find that the RNA velocity modelled by current methods of a target gene can be represented as a linear combination of TFs' expression level. On the scRNA-Seq pancreas dataset [4],given a target gene, we use Least Absolute Shrinkage and Selection Operator (LASSO) regression to predict its RNA velocity modelled by scVelo based on the expression level of candidate TFs and the target gene itself (**Supplementary Figure 1a**). For gene g, suppose that S_g represents the set of TFs with weights, $\boldsymbol{w_g}$, the RNA velocity of gene g can be written as:

$$v_g = \sum_{TF_i \in S_g} w_{g,TF_i} e_{TF_i} + w_{g,g} e_g \tag{1}$$

where e_{TF_i} and e_g are the expression level of TF_i and gene g respectively, w_{g,TF_i} and $w_{g,g}$ are entries of $\boldsymbol{w_g}$ corresponding to the weights of TF_i and the target gene itself.

Cells are separated into training and test sets to evaluate the LASSO model. The regression results on the test set shows a high correlation between the predicted and labeled velocity, as demonstrated in **Supplementary Figure 1 b-d**, which indicates that velocity of the target gene can be predicted by the expressions of its TFs. In addition, the TFs with non-zero positive weights are significantly enriched in TFs set functionally linked to the target gene (according to ENCODE TF-target datasets [21]) with p value of 2.89e-5 (one-sided t-test), further suggesting that velocity could be modelled by TFs. Moreover, due to regulatory relationship between TF and target, the expression levels of them change asynchronously, which could result in a clockwise curve on the phase portraits. And as expected, we find the desired clockwise curve on the phase portraits between TF-target pairs [20], which is similar to the theoretical prediction of existing velocity models on unspliced-spliced space (**supplementary figure 2**). These findings suggest that it is possible to construct a novel RNA velocity model based on the expression of TFs, instead of using unspliced/spliced RNA counts or any additional experimental labels.

Estimate RNA velocity with TFvelo.

Here we report TFvelo to estimate the RNA velocity of a target gene using the expression level of the target gene itself and its identified TFs. **Fig. 1a** provides an example on modeling the dynamics of gene LITAF, to explain why TFvelo outperforms existing methods (scVelo is used as the representative). Previous methods focus on modeling the expression level between the spliced and unspliced RNAs, while transcriptional dynamics of RNA splicing may not provide sufficient signal [18]. By comparison, TFvelo considers the relationship between the target gene and the overall regulation of its TFs, allowing for more robust and accurate modeling based on gene regulatory network.

For a target gene g, it does not model the dynamics of its unspliced/spliced RNA, instead, TFvelo explores the relationships between its expression level y_g and the regulation of TFs on g . In TFvelo, the RNA velocity $\frac{dy_g(t)}{dt}$ is determined by the expression level of involved TFs $\boldsymbol{X_g}$ and itself y_g,

$$\frac{dy_g(t)}{dt} = h\big(\boldsymbol{X_g}(t), y_g(t); \Psi_g\big). \tag{2}$$

Using a top-down strategy, which can relax the gene dynamics to more flexible profiles [8], TFvelo directly designs a profile function of target gene's expression level,

$$y_g(t) \ = f\big(t; \Phi_g\big). \tag{3}$$

where Ψ_g and Φ_g are two sets of gene-specific time-invariant parameters, which describe the influence of TFs on the target gene, and the shape of phase portraits, respectively. Our findings (**supplementary figure 1**) and early studies [22] have shown that linear models, which always have high interpretability and rarely suffer from over-fitting, can infer the regulation relationship by learned parameters. As a result, $h\big(\boldsymbol{X_g}(t), y_g(t); \Psi_g\big)$ is implemented with a linear model. The profile function of $y_g(t)$ can be chosen flexibly from a series of second-order differentiable functions [8]. For simplification, $y_g(t)$ is designed as a sine function (**Methods**), which is high-order differentiable, unimodal and able to model the dynamics without defining a switching time point between the periods of increasing expression and decreasing expression [4] [5] [9] [11] .

For optimization, all learnable parameters are divided into three groups, which are cell-specific latent time $t \in R^{n_{cell}}$, TFs' weights $\boldsymbol{W_g} \in R^{n_{TF}}$ and the shape parameters of phase portraits Ψ_g, in which n_{cell} and n_{TF} represent number of cells and number of genes, respectively. A generalized EM [23, 24] algorithm is adopted to minimize the loss function in each iteration, by updating the three groups of parameters alternately (**Fig. 1 g-i**). See **Methods** for the details.

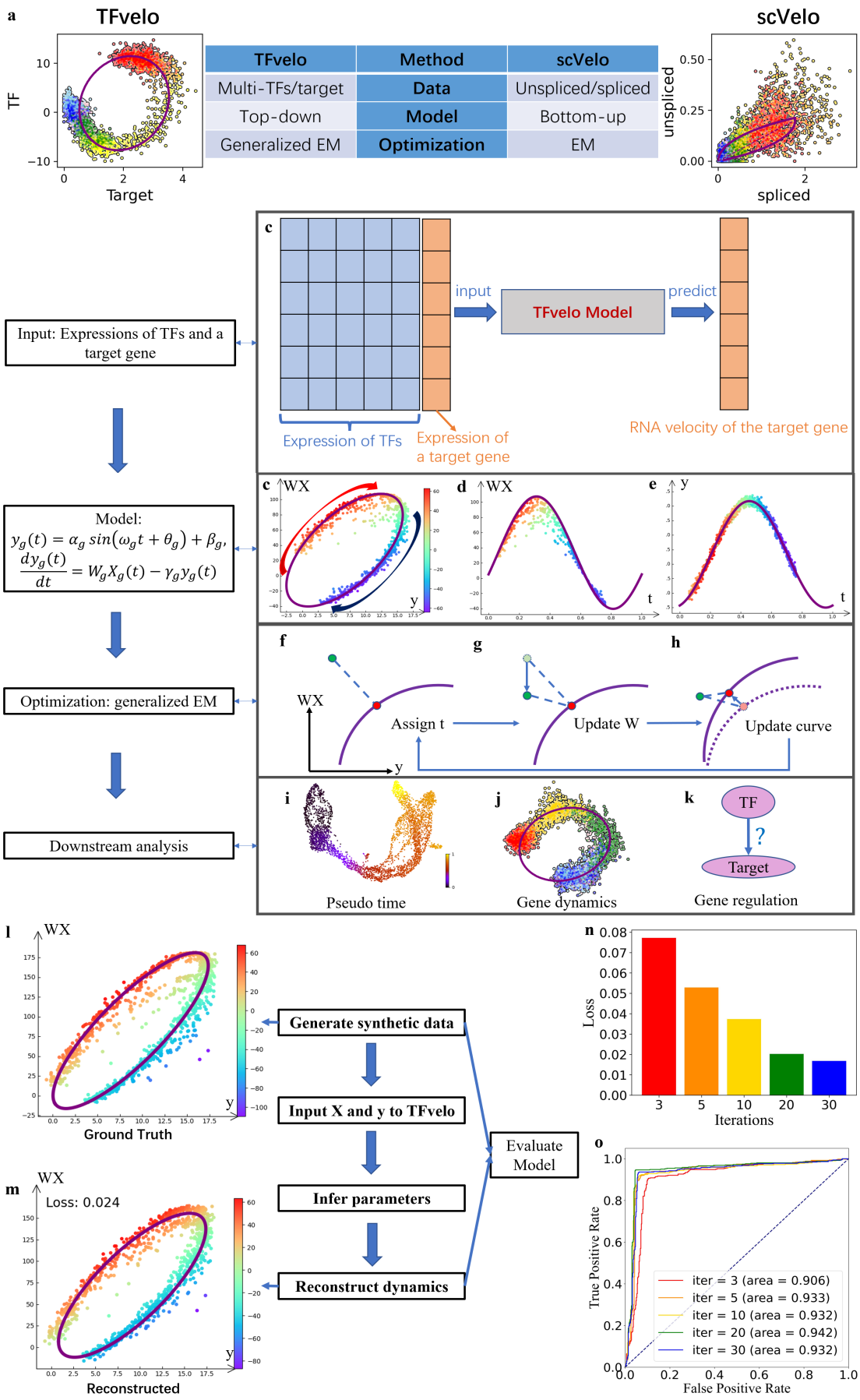

Figure 1. The overview of TFvelo. (a) The comparison between constructed dynamics

on LITAF gene in Pancreas scRNA-seq data by TFvelo and scVelo, respectively. (b) The workflow of TFvelo, where the RNA velocity is predicted based on the expression levels of TFs and the target gene. (c-e) The inferred joint phase plot of TF regulation and target gene, $(\boldsymbol{WX}, y)$, $\boldsymbol{WX}$ function w.r.t. Pseudo time, $(\boldsymbol{WX}, t)$ and y function w.r.t. pseudo-time(y, t) by TFvelo. The color bar indicates the RNA velocity in the plot, and the purple line refers to prediction of the theoretical model. For simplification, parameters used for modeling the dynamics of the gene g, $\boldsymbol{W}_g, \boldsymbol{X}_g, y_g$ and t_g are written as $\mathbf{W}, \mathbf{X}, \mathrm{y}$ and t, respectively. (f-h) The steps within each iteration of the generalized EM algorithm. Here the purple line represents the current prediction by the model. (f) Assignment of latent time to each cell (in green) to find the corresponding target point (shown in red), which is the nearest point to the cell on the purple line. (g) Optimization of the weights W to move all cells along the WX axis, minimizing the mean loss over all cells. (h) Optimization of the parameters in the dynamical function, to move the target point on the curve closer to each cell. (i-k) Downstream analysis of TFvelo, including pseudo-time inference, gene dynamics modeling, and key TF-target pair prediction. (l-o) Simulation on synthetic dataset. (l) An example of the generated synthetic dynamics. (m) Reconstructed dynamics by TFvelo of the sample shown in (l) with MSE loss. (n) The MSE loss b under different iterations on the synthetic dataset. (o) ROC curve under different iterations on TF-target pair prediction task.

TFvelo can reconstruct the dynamic model and detect the regulation relationship on synthetic datasets.

To evaluate TFvelo, we firstly verify it using a synthetic dataset with ground truth (see **Methods**), in which a target gene is regulated by 10 TFs, and the corresponding weights are either positive or negative, representing that TF may up-regulate or down-regulate the target gene, respectively. We randomly generate 100 synthetic gene dynamics with different ground truth parameters. For each of them, the TFvelo reconstructs the dynamic function and estimates the weight of each TF, based on the simulated expressions of TFs and target.

The performance of TFvelo was evaluated using normalized mean square error (MSE), F1 score, and area under the receiver operating characteristic (ROC) curve (**Supplementary Information part 3**). As demonstrated by the low loss (0.0203, **Fig. 1o**), the high F1 score (0.950) and the high area under the ROC curve (0.942, **Fig. 2p**), the TFvelo can effectively reconstruct the gene dynamics and detect the underlying gene regulation relationships.

TFvelo can model cell differentiation process on Pancreas dataset.

To evaluate TFvelo on scRNA-seq data, we first apply it to the dataset of E15.5 mouse pancreas, where the cell differentiation process can be detected by most splicing-based RNA velocity models. The pseudo-time predicted by TFvelo can identify the cell trajectory (**Fig. 2a**) well, and the arrows are drawn to reflect the direction of cells

development (**Fig. 2b, Methods**) using the derivative of pseudo-time. Although previous methods can generate a similar pseudo-time assignment and velocity stream (**Fig. 2c**), TFvelo can even achieve a higher velocity confidence than scVelo without using splicing information (**Fig. 2f**), which indicates a higher consistency of RNA velocities across neighboring cells in the UMAP space [5]. In addition, compared to scVelo, TFvelo's result on Epsilon and Delta cells can better align with the cell differentiation process (**Fig. 2c**). Next, we analyze the gene expression dynamics resolved along the pseudo-time of the top 300 likelihood-ranked genes, revealing a clear cascade of transcription (**Fig. 2d**). We then select genes that are highly expressed in cells with the earliest pseudo-time, corresponding to the stage of ductal progenitors. The following GO enrichment results (**Fig. 2e**) suggest that, consistent with conclusion by scVelo [5], the cell cycle process plays a critical role in ductal progenitors (cells within the purple cycle in **Fig. 2b**) .

As can be seen in **Figs. 2g-k** which illustrate five gene examples with high likelihoods, the theoretical dynamics of TFvelo can fit the joint distribution between transcription factors (TFs) and t target gene very well. According to the predicted dynamics by TFvelo, the expression level of SURF4 arrives at the highest value in beta cells, which is consist with the finding that SURF4 plays an important role in pancreatic beta cells [25]. By comparison, scVelo fails to detect any dynamics because of the insufficient information provided by unspliced mRNA (**Fig. 2g**). ANXA4 has been reported to be broadly expressed in the developing pancreas, and concentrated in the hepatopancreatic ducts and pancreatic islets, like beta cells [26]. TFvelo finds that the expression level of ANXA4 increases rapidly at the status of differentiating into pancreatic islets, and remains at a relative high level in beta cells (**Fig. 2h**). The expression of H19, which plays an important role in early development [27], is found to increase at the beginning and then decrease at the stage of Ngn3 high EP by TFvelo. By comparison, scVelo cannot distinguish cells from multiple states on H19 (**Fig. 2i**). LITAF, a gene highly expressed in ductal cells and lowly expressed in alpha, beta, gamma and delta cells [28], is also well modelled by TFvelo, while splice-based method captures a wrong dynamics for it (**Fig. 2j**). ECE1 which is validated to increase within the beta cell branch during pancreas development [29], is also found to rise from the beginning and reach its peak at beta cells in the prediction of TFvelo (**Fig. 2k**). See **supplementary Figures 4-5** for the dynamic models constructed by TFvelo on more genes, in which the joint plots on TFs-target space can form clockwise curves. These results suggest that TFvelo can provide more reasonable, robust and accurate dynamics for genes from single cell data.

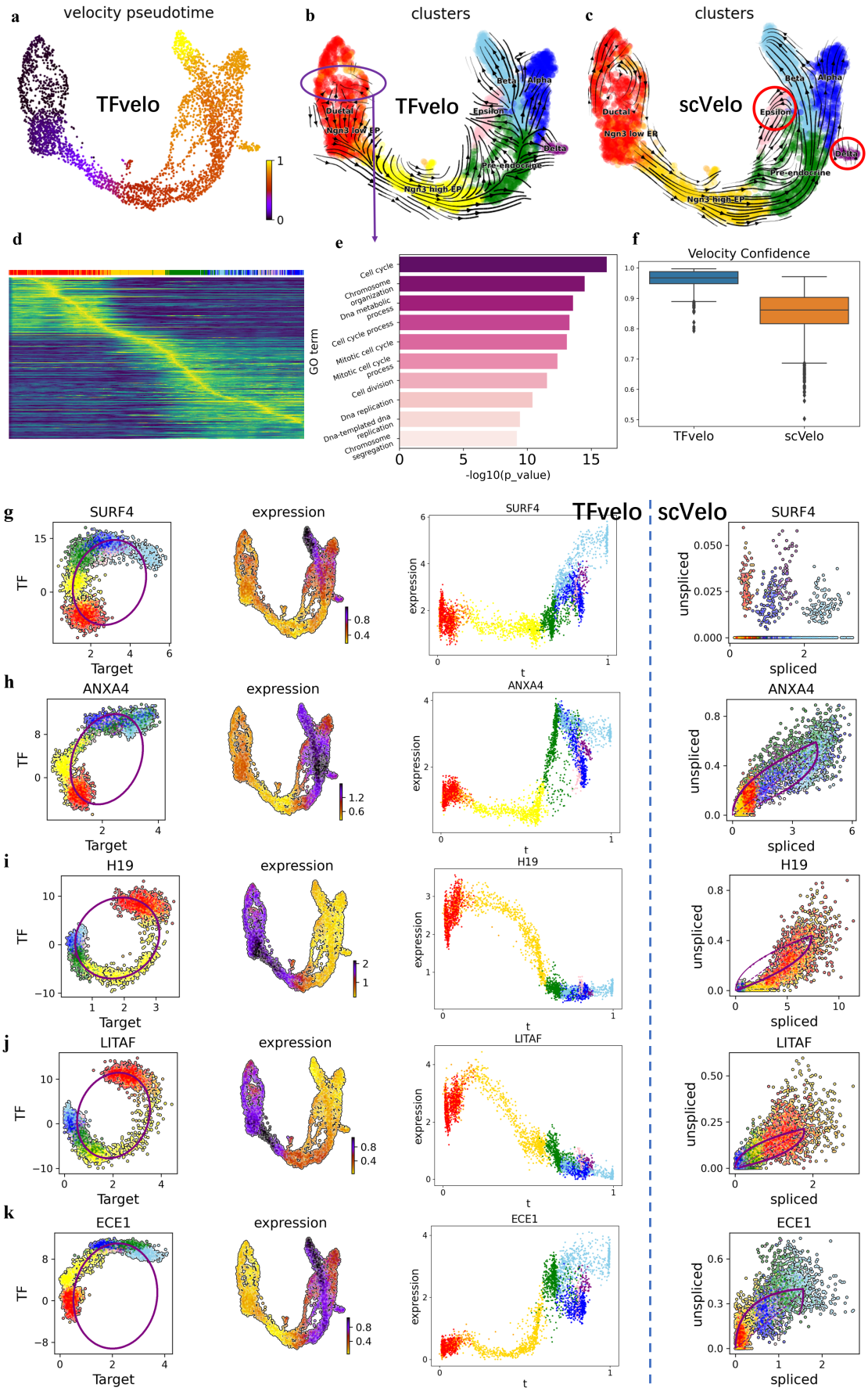

Figure 2. Results of TFvelo on pancreas dataset. (a) Pseudo-time inferred by TFvelo

in the UMAP-based embedding space. (b) The stream derived from the pseudo-time, in which each cell type is marked with a specific color. The Ngn3 low/high EP denotes the neurogenin3 low/high epithelial cells. (c) The stream provided by scVelo where the two red circles point out the non-smooth part learned by scVelo. (d) The dynamics heatmap of gene expression level resolved along pseudo-time for top 300 likelihood-ranked genes. (e) The top enriched GO terms for genes that are highly expressed in the cell in earliest pseudo-time. (f) The box plot of velocity confidence obtained by TFvelo and scVelo in (b) and (c). (g-k) The constructed dynamic model on five example genes with high likelihoods. In each row, the first three panels are results by TFvelo, including the dynamic phase plot of TF-target, target expression distribution on UMAP, and the target dynamics expression resolved along with pseudo-time, respectively. The last panel shows the dynamic model by scVelo on unspliced-spliced space.

What is more, TFvelo can also be used to detect key TF-target regulations. The insight behind TFvelo that the clockwise curve on the joint plot between two variables indicates the potential causality with time-delay, can provide a new perspective to infer the regulation relationship from single cell data. As illustrated in **Supplementary Figure 2,** there are many overlaps between those TF-target pairs associated with high weights and those TF-target pairs whose co-expression can form clockwise curves. Especially, JUND are assigned with high weights on the models of multiple target genes, and also tend to form clockwise curve (**Supplementary Figure 2c**). Early studies have reported that JUND modulates cell differentiation, proliferation, and apoptosis [30], further confirming the regulation mechanisms of JUND in pancreatic cells [31, 32].

TFvelo can model cell differentiation process on Gastrulation erythroid dataset. The Gastrulation Erythroid dataset is the transcriptional profile for mouse embryos [33]. Using TFvelo, the pseudo-time and cell development flow visualized by streamlines are consistent with the development in the erythroid lineage, from blood progenitors to erythroid cells (**Fig. 3a** and **Fig. 3b**). By comparison, scVelo cannot capture such developmental dynamics, as shown in **Fig. 3d** and **Fig. 3e**. The boxplot in **Fig. 3f** compares the velocity confidence obtained by TFvelo and scVelo, revealing that TFvelo achieves higher consistency of RNA velocities across neighboring cells on the UMAP.

Next three genes are selected as the examples to demonstrate that TFvelo results can be verified by solid biological findings from previous studies. Firstly, TACC1 is annotated as a gene involved in promoting cell division prior to the formation of differentiated tissues [34]. In our experiment, the expression of TACC1 initially increases, reaching its highest value in blood progenitors 2, which is just before differentiation into erythroid cells. Subsequently, TACC1's expression decreases with pseudo-time from erythroid cells 1 to erythroid cells 3 (**Fig. 3g**). Secondly, the expression of PFN1 decreases during the differentiation of blood cells from blood progenitors to mature erythroid cells (**Fig. 3h**). While previous research has reported that PFN1 is essential for the retention and metabolism of mouse hematopoietic stem cells in bone marrow

[35]. Thirdly, the expression level of TMEM14C increases during differentiation and reach high value in erythroid cells (**Fig. 3i**), which is consistent with the early findings that TMEM14C is required for erythroid mitochondrial heme metabolism [36]. By comparison, scVelo fails to detect any dynamics or detects wrong dynamics on these three genes. See **supplementary Figures 6-7** for the dynamic models constructed by TFvelo on more genes.

To detect the key TF-target regulation from the TFvelo results, we analyze both of the co-expression plot between TF-target pairs, and the weights associated with them. **Fig. 3j** shows that most TF-target pairs are associated with wights less than 1. We take two TF-target pairs with high wight as example for further analysis, which are FLI1-DIAPH1 (weight=1.23), and GATA2-CALR (weight=3.32). Firstly, these co-expression plots can form clockwise curves (**Fig. 3k**). Additionally, previous studies have reported these regulation pairs. For instance, FLI1 is known to regulate the actin assembly activity of diaphanous-related formins, including DIAPH1 [37]. In addition, GATA2 is verified to regulates hematopoietic stem cell proliferation and differentiation [38], and the CALR promoter contains regulatory sites for GATA [39]. . These findings demonstrate the potential of TFvelo in identifying key regulatory relationships from single cell data.

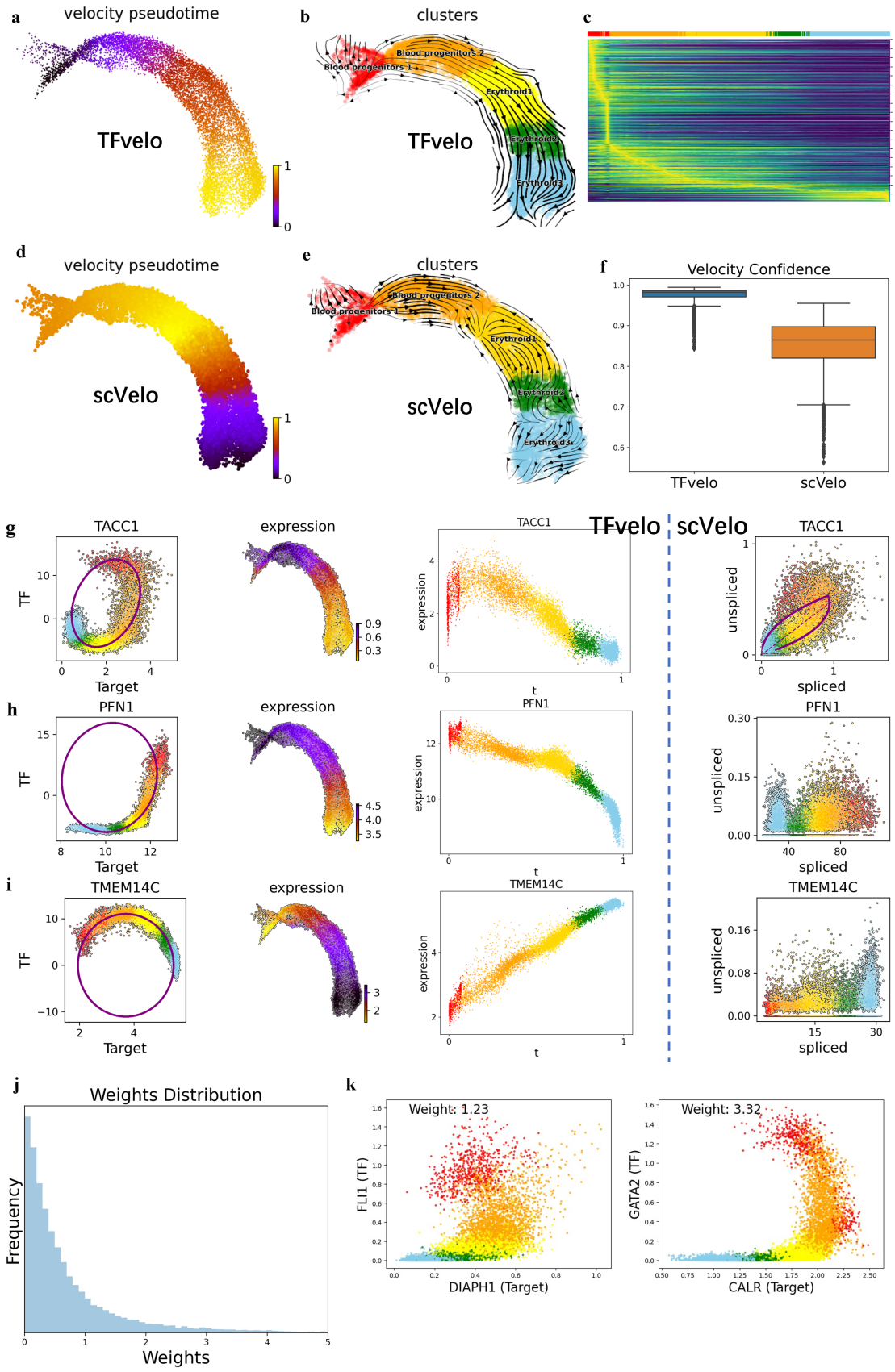

Figure 3. Results of TFvelo on gastrulation erythroid dataset. (a) Pseudo-time inferenced by TFvelo projected into a UMAP-based embedding. (b) The stream derived from the pseudo-time where each cell type is marked with a specific color. (c) The

heatmap of the dynamics of gene expression resolved along pseudo-time in the top 300 likelihood-ranked genes. (d) Pseudo-time inferenced by scVelo is projected into a UMAP-based embedding. (e) The stream provided by scVelo is projected into a UMAP-based embedding. (f) The box plot of velocity confidence obtained by TFvelo and scVelo. (g-i) The constructed dynamic model on three example genes with high likelihoods, which are TACC1, PFN1 and TMEM14C, respectively. From the left to the right within each row, the first three panels show results of TFvelo, including the dynamics on TFs-target space, expression distribution on UMAP, and the dynamics of expression resolved along with pseudo-time obtained, respectively. The last panel shows the constructed dynamic model of scVelo on unpliced-spliced space. (j) The distribution of weights that associated with all TF-target pairs. (k) Co-expression plot of 2 example TF-target pairs with high wight, which are FLI1-DIAPH1 (weight=1.23) and GATA2-CALR (weight=3.32).

TFvelo can fit the cell development process on human preimplantation embryos dataset.

Next TFvelo was applied to another single-cell RNA-seq dataset comprising 1,529 cells obtained from 88 human preimplantation embryos ranging from embryonic day 3 to 7 [40]. Since the raw sequencing file is not provided on the online dataset server (see **Data Availability Statement**), those RNA velocity methods relying on the spliced/unspliced information are not feasible. By contrast, TFvelo still successfully identifies the pseudo-time and developmental trajectory of the cells, as depicted in **Fig. 4a** and **Fig. 4b**. Additionally, dynamic models successfully fit the expression patterns of target genes, which are illustrated in **Fig.s 4d-f** as examples. In details, the dynamics predicted on PPP1R14A reveals a clockwise curve on TFs-target plot (**Fig. 4f**), and finds the changing point between the increasing and decreasing of expression level at day 5 post-fertilization. Consistent with the TFvelo results, it has been reported that the expression of PPP1R14A can be detected with high dynamics that it increases at the beginning and declines throughout the development in the early stage of zebrafish embryos [41]. TFvelo finds that the expression of RGS2 is the highest at the beginning, and decreases with the pseudo-time (**Fig. 4e**)which can be supported by the early findings that RGS2 plays a critical role in early embryo development [42]. In addition, GSTP1 was reported to exhibit oscillating expression changes in zebrafish, whose expression can be detected at a low level at the beginning of embryogenesis, then rises and reaches its maximum, followed by a sudden drop latter [43]. Consistent with this pattern, the expression of GSTP1 in human embryos reconstructed by TFvelo reaches its maximum at day 5 post-fertilization (**Fig. 4f**), and suddenly decreases after that time point. See **supplementary Figures 8-9** for the dynamic models constructed by TFvelo on more genes.

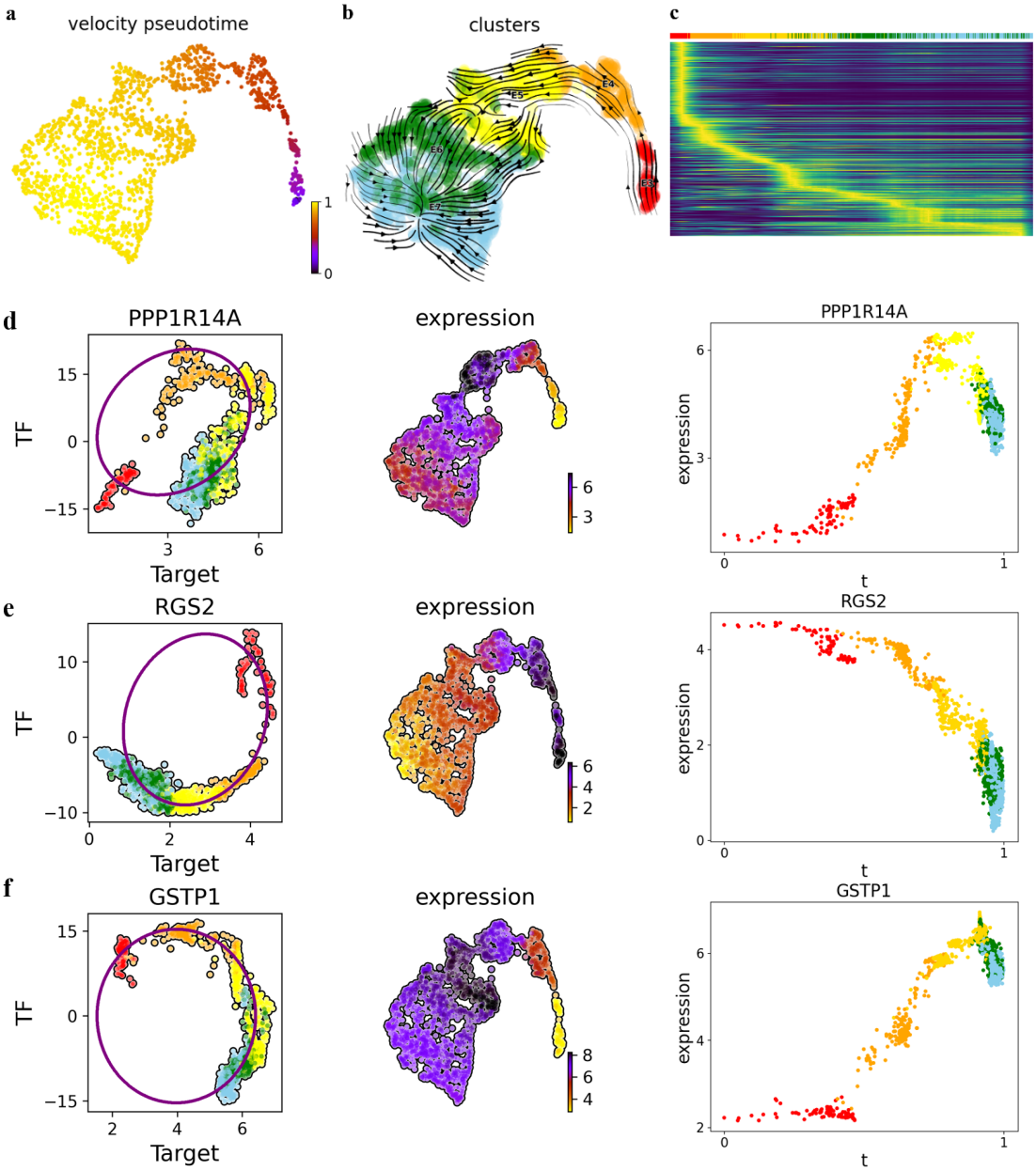

Figure 4. Results of TFvelo on human preimplantation embryos dataset. (a) Pseudo-time inferenced by TFvelo projected into a UMAP-based embedding. (b) The stream derived from the pseudo-time, where each cell type is marked with a specific color. (c) The heatmap which shows the dynamics of gene expression resolved along pseudo-time in the top 300 likelihood-ranked genes. (d-f) The constructed dynamic models on three example genes with high likelihoods, which are SERTAD1, RGS2 and GSTP1, respectively. From the left to the right within each row, the three subfigures show results of TFvelo, including the dynamics on TFs-target space, expression distribution on UMAP, and the dynamics of expression resolved along with pseudo-time obtained, respectively.

Discussion

The recent development in single-cell RNA velocity approaches provide opportunities to detect cell trajectory on single cell datasets. However, the existing methods often fails to fit the cell dynamics well in the unspliced/spliced space, which is because of the weak signal and high noise at single cell resolution. Motivated by the insight that the RNA velocity could be driven by the transcriptional regulation, we propose TFvelo from a new avenue, which can estimate the RNA velocity based on gene regulatory network. Different from previous approaches which require the splicing information from the data, TFvelo models the dynamics of a target gene with the expression level of TFs and the target gene itself. In the computation framework, a generalized EM algorithm is adopted, so that the latent-time, the weight of TFs and those learnable parameters in the dynamic function can be updated alternately in each iteration.

Experiments on a synthetic dataset validate the effectiveness of TFvelo. Results on various scRNA-seq datasets further demonstrate that TFvelo can fit the gene dynamics well, showing a desired clockwise curve on the phase plot between TFs and the target. Furthermore, TFvelo can be used to infer the pseudo time, cell trajectory and detect key TF-target regulation. Compared with the previous RNA velocity model relying on splicing information, TFvelo can achieve better performance on those scRNA-seq datasets, and be flexibly applied to more categories of datasets without splicing information.

In conclusion, this study provides a novel idea to construct RNA dynamics model based on gene regulatory relationship. In the future, more complex models may be explored to replace the current linear model in TFvelo for estimating the weight of each TFs. The employed sine function for describing expression dynamics could also be replaced by other functional forms flexibly. In addition, those recently proposed ideas for improving unspliced/spliced-based RNA velocity models can also be adopted in the further development of TFvelo, including stochastics modeling with Bayesian inference [9], multi-omics integration [15] and universe time inference [8].

Methods

Data preprocessing.

We utilize the data preparation procedure with the default setting in scVelo [5], where the difference is that TFvelo is applied to the total mRNA counts, which is the sum of unspliced and spliced counts. Additionally, after filtering out those genes that can be detected from fewer than 2% of cells, and selecting the top 2,000 highly variable genes

(HVGs), we normalize the expression profile of these HVGs by the total count in each cell. A nearest-neighbor graph (with 30 neighbors by default) was calculated based on Euclidean distances in principal component analysis space (with 30 principal components by default) on logarithmized spliced counts. After that, first-order and second-order moments (means and uncentered variances) are computed for each cell across its nearest neighbors. These steps are implemented with scv.pp.filter_and_normalize() and scv.pp.moments().

Additionally, to find potential transcriptional relationship within the selected 2,000 highly variable genes, we refer to the TF-target pairs annotated in ENCODE TF-target dataset [21] and ChEA TF-target dataset [44]. If the regulatory relationship between a TF-target pair is labeled in at least one of these two datasets, the TF will be included in the involved TFs set when modeling the dynamic of the target gene.

The computational framework of TFvelo

Given a gene g, TFvelo assumes that the RNA velocity $\frac{dy_g(t)}{dt}$ is determined by the expression level of involved TFs $\boldsymbol{X_g}$ and the target gene itself y_g, which is $\frac{dy_g(t)}{dt} = h\left(\boldsymbol{X_g}(t), y_g(t); \Psi_g\right)$. Using a top-down strategy [8], TFvelo directly designs a profile function of target gene's expression level to relax the gene dynamics to more flexible profiles, that is $y_g(t) = f\left(t; \Phi_g\right)$. Ψ_g and Φ_g are two sets of gene-specific, time-invariant parameters, which control the influence of TFs on the target gene, and the shape of target gene's phase portraits, respectively. Our findings and early studies [22] have shown that it is feasible to infer the regulation relationship according to the learned parameter in linear regression models based on expression data. Additionally, linear models always have high interpretability, and rarely suffer from over-fitting. As a result, to show the capability of modeling RNA velocity based on gene regulatory relationship, we simply adopt a linear model as $h\left(\boldsymbol{X_g}(t), y_g(t); \Psi_g\right)$, which can map the expression levels of multiple TFs $\boldsymbol{X_g} \in R^{n_{TF}}$ from n_{TF}-dimensional space to 1-dimensional representation $\boldsymbol{W_g}\boldsymbol{X_g}$ with a weight vector $\boldsymbol{W_g}$, where n_{TF} represents the number of involved TFs. As a result, the dynamic equation between TFs and the target gene g can be written as

$$\frac{dy_g(t)}{dt} = h\left(\boldsymbol{X_g}(t), y_g(t); \Psi_g\right) = \boldsymbol{W_g}\boldsymbol{X_g}(t) - \gamma_g y_g(t)\,, \tag{4}$$

where $t \in [0,1)$ is the latent time assigned to each cell, γ_g is the degradation rate.

Besides, The profile function of $y_g(t)$ can be chosen flexibly from those second-order differentiable functions [8]. For simplification, $y_g(t)$ is designed with a sine function

in our implementation, which is high-order differentiable and suitable for capture the curves on phase plot.

$$y_g(t) = f(t; \Phi_g) = \alpha_g \, sin(\omega_g t + \theta_g) + \beta_g, \tag{5}$$

where $\alpha_g, \beta_g, \theta_g$ are gene-specific parameters to be learned. ω_g is fixed as 2π, so that the $y_g(t)$ is unimodal within $t \in [0,1)$, which follows the common assumption shared by most current approaches. By employing the sin function, TFvelo can model the dynamics of each gene without defining a switching time point between the of increasing and decreasing periods [4] [5] [9] [11]. To optimize the parameters in TFvelo, a generalized EM [23, 24] algorithm is adopted, as discussed in the following.

Optimization of generalized EM algorithm

To simplify the mathematical representation, here we denote the $\boldsymbol{W}_g, \boldsymbol{X}_g, y_g, \alpha_g, \beta_g, \theta_g, \gamma_g$ and ω_g as $\boldsymbol{W}, \boldsymbol{X}, y, \alpha, \beta, \theta, \gamma$ and ω, respectively. As a result, the equation (5) and (4) can be written as

$$y(t) = \alpha \sin(2\pi t + \theta) + \beta, \tag{6}$$

$$\frac{dy(t)}{dt} = \boldsymbol{W}\boldsymbol{X}(t) - \gamma y. \tag{7}$$

From the dynamics model in equation (6) and (7), we can derive that,

$$\boldsymbol{W}\boldsymbol{X}(t) = \alpha\sqrt{4\pi^2 + \gamma^2}\, sin(2\pi t + \theta + \phi) + \beta\gamma, \phi = arctan\left(\frac{2\pi}{\gamma}\right), \tag{8}$$

For modeling a gene, suppose observations $\boldsymbol{X}_c^{obs}$ and y_c^{obs} be the normalized expression levels of TFs and the target gene within a cell c, s_c^{obs} be the state space representation of the cell c, and $\hat{s}(t_c)$ be the modeled state at time t_c, we have $s_c^{obs} = [\boldsymbol{W}\boldsymbol{X}_c^{obs}, y_c^{obs}]$ and $\hat{s}(t_c) = [\boldsymbol{W}\boldsymbol{X}(t_c), y(t_c)]$. Aiming at finding the phase trajectory $\hat{s}(t)$ that best fits the observations for all cells, we define the loss function as:

$$L = \sum_c ||s_c^{obs} - \hat{s}(t_c)||^2 = \sum_c (\boldsymbol{W}\boldsymbol{X}_c^{obs} - \boldsymbol{W}\boldsymbol{X}(t_c))^2 + (y_c^{obs} - y(t_c))^2, \tag{9}$$

The residuals of the observations are defined as $e_c = sign(y_c^{obs} - y(t_c))||s_c^{obs} - \hat{s}(t_c)||^2$, which is assumed to follow a normal distribution $e_c \sim N(0, \sigma^2)$, where the gene-specific σ is constant across cells, and the observations are independent and identically distributed. Then we can arrive at the likelihood function,

$$\mathcal{L}(W, \alpha, \beta, \theta, \gamma) = \frac{1}{\sqrt{2\pi}\sigma} * \exp\left(-\frac{1}{2\sigma^2}\sum_c ||s_c^{obs}(\boldsymbol{W}) - \hat{s}_{t_c}(\boldsymbol{W}, \alpha, \beta, \theta, \gamma)||^2\right), \tag{10}$$

Where $\hat{s}_{t_c}(\boldsymbol{W}, \alpha, \beta, \theta, \gamma)$ represents $\hat{s}(t_c; \boldsymbol{W}, \alpha, \beta, \theta, \gamma)$. Finally, the optimization algorithm aims to minimize the negative log-likelihood function:

$$l(\boldsymbol{W}, \alpha, \beta, \theta, \gamma) = -\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) + \frac{1}{2\sigma^2}\sum_c ||s_c^{obs}(\boldsymbol{W}) - \hat{s}_{t_c}(\boldsymbol{W}, \alpha, \beta, \theta, \gamma)||^2, \tag{11}$$

In standard EM algorithms, the E-step computes the expected value of likelihood function given the observed data and the current estimated parameters to update the latent variables, and the M-step consists of maximizing that expectation computed in E step by optimizing the parameters. While under some circumstances, there is an intractable problem in M step, which could be addressed by generalized EM algorithms. Expectation conditional maximization is one form of generalized EM algorithms, which makes multiple constraint optimizations within each M step [23]. In detail, the parameters can be separated into more subsets, and the M step consists of multiple steps as well, each of which optimizes one subset of parameters with the reminder fixed [24].

In TFvelo, all learnable parameters can be divided into three groups, which are cell-specific latent time $t \in R^{n_{cell}}$, weight of each TF $\mathrm{W} \in R^{n_{TF}}$ and scalars $[\alpha, \beta, \theta, \gamma]$ respectively, where n_{cell} and n_{TF} represent the number of cells and the number of involved TFs. Generalized EM algorithm is adopted so that the three group of parameters will be updated alternately to minimize the loss function in each iteration:

a) Assign t within [0,1) for each cell by grid search. For each cell, on the phase plot, the closest point located on the curve described by the dynamic model is defined as the target point. The step b) and c) aim to minimize the average distance between cells and their corresponding target points.
b) Update $\boldsymbol{W}$ by linear regression with bounds on these weights to minimize the loss function. Trust region reflective algorithm [45], which is an interior-point-like method is adopted to solve this linear least-squares problem with inequality constraint. In our implementation, the weights are bounded within (-20, 20) by default.
c) Update $[\alpha, \beta, \theta, \gamma]$ with constraints of $\alpha \in (0, \infty)$, $\theta \in (-\pi, \pi)$ and $\gamma \in (0, \infty)$ to minimize the loss function.

By iteratively running the three steps, parameters of TFvelo for each target gene can be optimized. Furthermore, multiple downstream analysis could be conducted by combining the learned dynamics of all target genes.

Initialization of the generalized EM algorithm.

Since EM algorithms could be trapped in local minimum, parameters are initialized in the following steps. Firstly, those cells in which expression of target gene is 0 are filtered out. Then X, y are normalized by $X = X/var(X)$, $y = y/var(y)$, where $var(\cdot)$ represents the standard deviation function. The weights W is initialized with the value of spearman correlation of each TF-target pair. For parameters α, β and γ, according to the steady-state assumption that $\dot{y}|_{y=max\,(y)} = 0$, γ can be initialized by $\gamma = \frac{W\mathrm{X}}{y}|_{y \to y_{max}}$. At the points of maximal and minimal y value, we have $y_{\max} = \alpha + \beta$ and $\mathrm{y}_{\min} = -\alpha + \beta$. As a result, α and β can be initialized by $\alpha = \frac{y_{max} - y_{min}}{2}$ and

$\beta = \frac{y_{max}+y_{min}}{2}$, respectively. These assumptions are only used for initialization and in the following optimization process, all parameters will be updated iteratively. The hyper-parameter ω is set as 2π and parameter θ isintialized with various values, so that the generalized EM algorithm will run at different start points. After that, the model with the lowest loss is selected.

Inference of Pseudo-time

After constructing the dynamic model for each gene, TFvelo will infer a gene-agnostic pseudo-time for each cell based on the combination of RNA velocities of all genes, in the same way with scVelo [5]. For this purpose, we first compute the cosine similarities between velocities and potential cell state transitions to get the transition matrix. Then the end points and root cells are obtained as stationary states of the velocity-inferred transition matrix and its transposed, respectively. After inferring the root cells' location on the embedding space, velocity pseudo-time, which measures the average number of steps it takes to reach a cell after start walking from root cells, can be computed. The computation of pseudo-time can be done by using the implementation of scVelo by scv.tl.velocity_pseudotime().

Velocity streams on UMAP

To improve the r consistency of the trajectory inferred by Pseudo-time and stream, we draw the stream based on the derivative of pseudo-time, instead of transition matrix []. The 2D UMAP space is clipped into 32×32 blocks by default. The pseudo-time of each block is defined as the average pseudo-time of all cells within this block. Then the derivative of pseudo-time over these blocks is calculated to draw streams. To smooth these streams, the flow on each block is refined by those of its neighboring blocks.

Heatmap and GO term analysis

The gene expressions along with pseudo-time can be plotted as a heatmap by scv.pl.heatmap(). For pancreas dataset, genes whose expression peak is located within the earliest 10% of pseudo-time are selected as highly expressed genes of ductal progenitors. With the whole gene list in the dataset as the background, we carry out GO term enrichment analysis for ductal progenitors to construct a functional enrichment profile. According to the GO analysis server (http://geneontology.org/), the statistical significance is obtained according to Fisher's exact test by default, and GO terms with False Discovery Rate (FDR) smaller than 0.05 are shown in **Fig. 2e**.

Data Availability Statement

The pancreatic endocrinogenesis dataset is available from NCBI GEO under accession ID GSE132188. It comprises the single-cell RNA-seq (10X) data of 27,998 genes of 3,696 pancreatic epithelial and Ngn3-Venus fusion cells sampled from mouse embryonic day 15.5. Processed data is acquired from scv.datasets.pancreas().

The gastrulation erythroid dataset, which is selected from the transcriptional profiles of mouse embryos [33], provides expressions of 53,801 genes of 9,815 cells. This dataset is incorporated by scv.datasets.gastrulation_erythroid().

Human preimplantation embryos dataset is a single-cell RNA-seq dataset of 1,529 cells obtained from 88 human preimplantation embryos ranging from embryonic day 3 to 7. Data were downloaded with the accession number of E-MTAB-3929 from EMBL-EBI [40]. In this dataset, only the RNA abundance is provided. As a result, those RNA velocity methods relying on the spliced/unspliced are not available.

Code Availability Statement

TFvelo is implemented in Python, based on the scVelo package [5]. The source code can be downloaded from both of the supporting website, http://www.csbio.sjtu.edu.cn/, and the Github repository, https://github.com/xiaoyeye/TFvelo.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 61725302, 62073219 and 62103262), Shanghai Pujiang Program (No. 21PJ1407700), and the Science and Technology Commission of Shanghai Municipality (22511104100).

References

1. Saliba A-E, Westermann AJ, Gorski SA, Vogel J. Single-cell RNA-seq: advances and future challenges. Nucleic acids research. 2014;42(14):8845-60.
2. Chen G, Ning B, Shi T. Single-cell RNA-seq technologies and related computational data analysis. Frontiers in genetics. 2019:317.
3. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nature Reviews Genetics. 2019;20(5):273-82.
4. La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, et al. RNA velocity of single cells. Nature. 2018;560(7719):494-8.
5. Bergen V, Lange M, Peidli S, Wolf FA, Theis FJ. Generalizing RNA velocity to transient cell states through dynamical modeling. Nature biotechnology. 2020;38(12):1408-14.
6. Qiao C, Huang Y. Representation learning of RNA velocity reveals robust cell transitions.

Proceedings of the National Academy of Sciences. 2021;118(49):e2105859118.
7. Riba A, Oravecz A, Durik M, Jiménez S, Alunni V, Cerciat M, et al. Cell cycle gene regulation dynamics revealed by RNA velocity and deep-learning. Nature communications. 2022;13(1):1-13.
8. Gao M, Qiao C, Huang Y. UniTVelo: temporally unified RNA velocity reinforces single-cell trajectory inference. Nature Communications. 2022;13(1):6586.
9. Qin Q, Bingham E, La Manno G, Langenau DM, Pinello L. Pyro-Velocity: Probabilistic RNA Velocity inference from single-cell data. bioRxiv. 2022.
10. Gu Y, Blaauw D, Welch JD. Bayesian inference of rna velocity from multi-lineage single-cell data. bioRxiv. 2022.
11. Qiu X, Zhang Y, Martin-Rufino JD, Weng C, Hosseinzadeh S, Yang D, et al. Mapping transcriptomic vector fields of single cells. Cell. 2022;185(4):690-711. e45.
12. Cao J, Zhou W, Steemers F, Trapnell C, Shendure J. Sci-fate characterizes the dynamics of gene expression in single cells. Nature biotechnology. 2020;38(8):980-8.
13. Ma A, McDermaid A, Xu J, Chang Y, Ma Q. Integrative methods and practical challenges for single-cell multi-omics. Trends in biotechnology. 2020;38(9):1007-22.
14. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. Bioinformatics and biology insights. 2020;14:1177932219899051.
15. Gorin G, Svensson V, Pachter L. Protein velocity and acceleration from single-cell multiomics experiments. Genome biology. 2020;21(1):1-6.
16. Li C, Virgilio M, Collins KL, Welch JD, editors. Single-cell multi-omic velocity infers dynamic and decoupled gene regulation. International Conference on Research in Computational Molecular Biology; 2022: Springer.
17. Liu R, Pisco AO, Braun E, Linnarsson S, Zou J. Dynamical Systems Model of RNA Velocity Improves Inference of Single-cell Trajectory, Pseudo-time and Gene Regulation. Journal of Molecular Biology. 2022:167606.
18. Bergen V, Soldatov RA, Kharchenko PV, Theis FJ. RNA velocity—current challenges and future perspectives. Molecular systems biology. 2021;17(8):e10282.
19. Gao M, Qiao C, Huang Y. UniTVelo: temporally unified RNA velocity reinforces single-cell trajectory inference. bioRxiv. 2022.
20. Yuan Y, Bar-Joseph Z. Deep learning of gene relationships from single cell time-course expression data. Briefings in bioinformatics. 2021;22(5).
21. Feingold E, Good P, Guyer M, Kamholz S, Liefer L, Wetterstrand K, et al. The ENCODE (ENCyclopedia of DNA elements) project. Science. 2004;306(5696):636-40.
22. Tran A, Yang P, Yang JY, Ormerod JT. scREMOTE: Using multimodal single cell data to predict regulatory gene relationships and to build a computational cell reprogramming model. NAR genomics and bioinformatics. 2022;4(1):lqac023.
23. Meng X-L, Rubin DB. Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika. 1993;80(2):267-78.
24. Bishop CM, Nasrabadi NM. Pattern recognition and machine learning: Springer; 2006.
25. Saegusa K, Matsunaga K, Maeda M, Saito K, Izumi T, Sato K. Cargo receptor Surf4 regulates endoplasmic reticulum export of proinsulin in pancreatic β-cells. Communications Biology. 2022;5(1):458.

26. Zhang D, Golubkov VS, Han W, Correa RG, Zhou Y, Lee S, et al. Identification of Annexin A4 as a hepatopancreas factor involved in liver cell survival. Developmental biology. 2014;395(1):96-110.
27. Kim D, Zhang L, Dzau V, Pratt R. H19, a developmentally regulated gene, is reexpressed in rat vascular smooth muscle cells after injury. The Journal of clinical investigation. 1994;93(1):355-60.
28. Segerstolpe Å, Palasantza A, Eliasson P, Andersson E-M, Andréasson A-C, Sun X, et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. Cell metabolism. 2016;24(4):593-607.
29. Byrnes LE, Wong DM, Subramaniam M, Meyer NP, Gilchrist CL, Knox SM, et al. Lineage dynamics of murine pancreatic development at single-cell resolution. Nature communications. 2018;9(1):3922.
30. Hernandez J, Floyd D, Weilbaecher K, Green P, Boris-Lawrie K. Multiple facets of junD gene expression are atypical among AP-1 family members. Oncogene. 2008;27(35):4757-67.
31. Good AL, Cannon CE, Haemmerle MW, Yang J, Stanescu DE, Doliba NM, et al. JUND regulates pancreatic β cell survival during metabolic stress. Molecular metabolism. 2019;25:95-106.
32. Wang K, Cui Y, Lin P, Yao Z, Sun Y. JunD Regulates Pancreatic β-Cells Function by Altering Lipid Accumulation. Frontiers in Endocrinology. 2021:755.
33. Pijuan-Sala B, Griffiths JA, Guibentif C, Hiscock TW, Jawaid W, Calero-Nieto FJ, et al. A single-cell molecular map of mouse gastrulation and early organogenesis. Nature. 2019;566(7745):490-5.
34. Guyot R, Vincent S, Bertin J, Samarut J, Ravel-Chapuis P. The transforming acidic coiled coil (TACC1) protein modulates the transcriptional activity of the nuclear receptors TR and RAR. BMC Molecular Biology. 2010;11:1-13.
35. Zheng J, Lu Z, Kocabas F, Böttcher RT, Costell M, Kang X, et al. Profilin 1 is essential for retention and metabolism of mouse hematopoietic stem cells in bone marrow. Blood, The Journal of the American Society of Hematology. 2014;123(7):992-1001.
36. Yien YY, Robledo RF, Schultz IJ, Takahashi-Makise N, Gwynn B, Bauer DE, et al. TMEM14C is required for erythroid mitochondrial heme metabolism. The Journal of clinical investigation. 2014;124(10):4294-304.
37. Higashi T, Ikeda T, Murakami T, Shirakawa R, Kawato M, Okawa K, et al. Flightless-I (Fli-I) regulates the actin assembly activity of diaphanous-related formins (DRFs) Daam1 and mDia1 in cooperation with active Rho GTPase. Journal of Biological Chemistry. 2010;285(21):16231-8.
38. Nanaa A, Viswanatha D, Xie Z, Jevremovic D, Nguyen P, Salama ME, et al. Clinical and biological characteristics and prognostic impact of somatic GATA2 mutations in myeloid malignancies: a single institution experience. Blood cancer journal. 2021;11(6):122.
39. Qiu Y, Michalak M. Transcriptional control of the calreticulin gene in health and disease. The international journal of biochemistry & cell biology. 2009;41(3):531-8.
40. Chu L-F, Leng N, Zhang J, Hou Z, Mamott D, Vereide DT, et al. Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. Genome biology. 2016;17:1-20.
41. Lang I, Virk G, Zheng DC, Young J, Nguyen MJ, Amiri R, et al. The evolution of duplicated genes of the Cpi-17/Phi-1 (ppp1r14) family of protein phosphatase 1 inhibitors in teleosts.

International Journal of Molecular Sciences. 2020;21(16):5709.
42. Zhu Y, Jiang Y-H, He Y-P, Zhang X, Sun Z-G, Jiang M-X, et al. Knockdown of regulator of G-protein signalling 2 (Rgs2) leads to abnormal early mouse embryo development in vitro. Reproduction, Fertility and Development. 2015;27(3):557-66.
43. Hrubik J, Glisic B, Fa S, Pogrmic-Majkic K, Andric N. Erk-Creb pathway suppresses glutathione-S-transferase pi expression under basal and oxidative stress conditions in zebrafish embryos. Toxicology Letters. 2016;240(1):81-92.
44. Lachmann A, Xu H, Krishnan J, Berger SI, Mazloom AR, Ma'ayan A. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. Bioinformatics. 2010;26(19):2438-44.
45. Branch MA, Coleman TF, Li Y. A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. SIAM Journal on Scientific Computing. 1999;21(1):1-23.