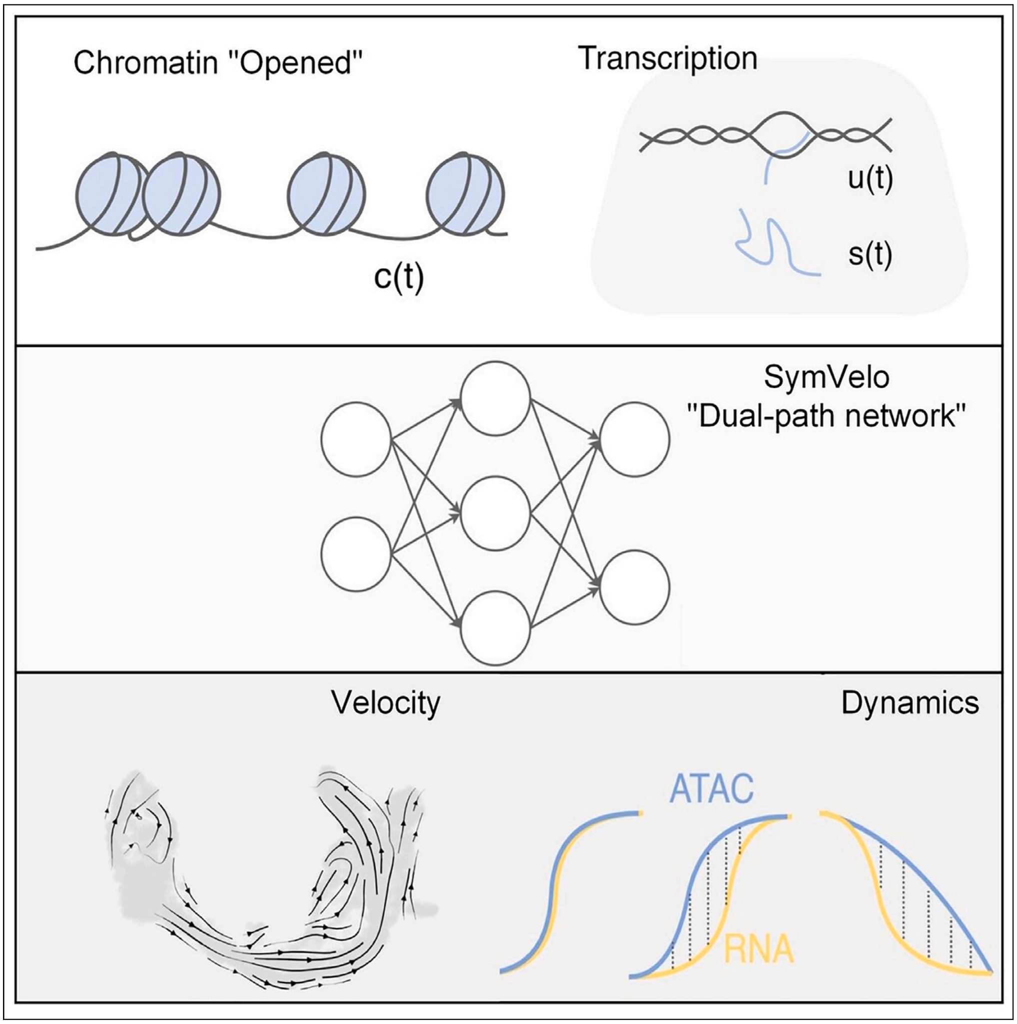
iScience

CellPress
OPEN ACCESS

Article

RNA velocity prediction via neural ordinary differential equation



Chenxi Xie, Yueyuxiao Yang, Hao Yu, ..., Bin Dong, Li Zhang, Meng Yang

zhangli_pku@pku.edu.cn (L.Z.)
yangmeng1@mgi-tech.com (M.Y.)

Highlights

Our study introduces a dual-path neural network framework for RNA velocity estimation

High-order polynomials adeptly depict intricate multi-lineage systems

SymVelo's RNA velocity estimation is smoother and aligns closely with actual cellular time

SymVelo can accommodate two or more modalities with ease

Xie et al., iScience *27*, 109635
April 19, 2024 © 2024 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.isci.2024.109635

iScience

CellPress OPEN ACCESS

Article

RNA velocity prediction via neural ordinary differential equation

Chenxi Xie,[1,3] Yueyuxiao Yang,[1,3] Hao Yu,[2,3] Qiushun He,[1] Mingze Yuan,[2] Bin Dong,[2] Li Zhang,[2,*] and Meng Yang[1,4,*]

SUMMARY

RNA velocity is a crucial tool for unraveling the trajectory of cellular responses. Several approaches, including ordinary differential equations and machine learning models, have been proposed to interpret velocity. However, the practicality of these methods is constrained by underlying assumptions. In this study, we introduce SymVelo, a dual-path framework that effectively integrates high- and low-dimensional information. Rigorous benchmarking and extensive studies demonstrate that SymVelo is capable of inferring differentiation trajectories in developing organs, analyzing gene responses to stimulation, and uncovering transcription dynamics. Moreover, the adaptable architecture of SymVelo enables customization to accommodate intricate data and diverse modalities in forthcoming research, thereby providing a promising avenue for advancing our understanding of cellular behavior.

INTRODUCTION

Methods for RNA velocity estimation can be broadly categorized into two types: traditional model-based approaches and data-driven methods. Traditional methods[1,2] use gene-specific first-order ordinary differential equations (ODEs) to model transcriptional dynamics and approximate kinetic parameters through extreme-quantile linear regression or expectation-maximization (EM) algorithm. Nonetheless, these approaches are susceptible to high noise levels, potentially yielding less reliable differentiation trajectories. Furthermore, these conventional model-based methods heavily rely on certain assumptions that only a subset of genes adhere to simple kinetics, thereby imposing limitations on their applicability.[3]

Conversely, data-driven methods leverage the power of machine learning to tackle the challenges posed by traditional methods.[4–8] VeloAE,[7] for example, utilizes a custom autoencoder to learn a low-dimensional representation of RNA velocity, effectively mitigating noise in high-dimensional count data via cellular state projection. Additionally, it accounts for inter-gene relationships via representation learning, which is often ignored in conventional gene-specific models. Despite its proficiency in accurately estimating cellular transitions, VeloAE falls short in revealing gene-level characteristics directly. This shortcoming stems from the integration of different gene information within a low-dimensional latent space in a "black box" manner, which lacks biological interpretability. Furthermore, the supervision of low-dimensional representations during training is still based on the steady-state mode, which only restricts the extreme-quantile cells for each dimension within the latent space.

To address this shortcoming, we propose a dual-path framework to estimate RNA velocity, which simultaneously trains two branches of neural networks to handle high- and low-dimensional RNA velocities. The framework aligns these branches via mutual learning to inherit the robustness of representation learning from the low-dimensional branch while preserving biological interpretability through the high-dimensional counterpart. Moreover, mutual learning covers all cells for each latent dimension, providing inter-gene information in the supervision of representation learning.

Our proposed approach, SymVelo, is validated across various developmental trajectories in the dentate gyrus, single-cell metabolically labeled new RNA tagging sequencing (scNT-seq), and multi-modality dataset. It elucidates the direction of differentiation in the hippocampal dentate gyrus during neurogenesis, retrieves response gene patterns of neuronal activity, and identifies distinctive transcription dynamics in multi-modality with high complexity and sparseness. Our results suggest that SymVelo represents a promising approach to estimate RNA velocity while preserving biological interpretability, which is critical for advancing our understanding of complex biological processes.

RESULTS

Profile and robust performance of SymVelo

We introduce SymVelo, a comprehensive framework for the prediction of RNA velocity and the analysis of cell transitions. As depicted in Figure 1, SymVelo consists of three modules: the temporal difference module, the pre-trained representation learning module, and the mutual

[1] MGI, BGI-Shenzhen, Shenzhen 518083, China
[2] Peking University, Beijing 100871, China
[3] These authors contributed equally
[4] Lead contact
*Correspondence: zhangli_pku@pku.edu.cn (L.Z.), yangmeng1@mgi-tech.com (M.Y.)

https://doi.org/10.1016/j.isci.2024.109635

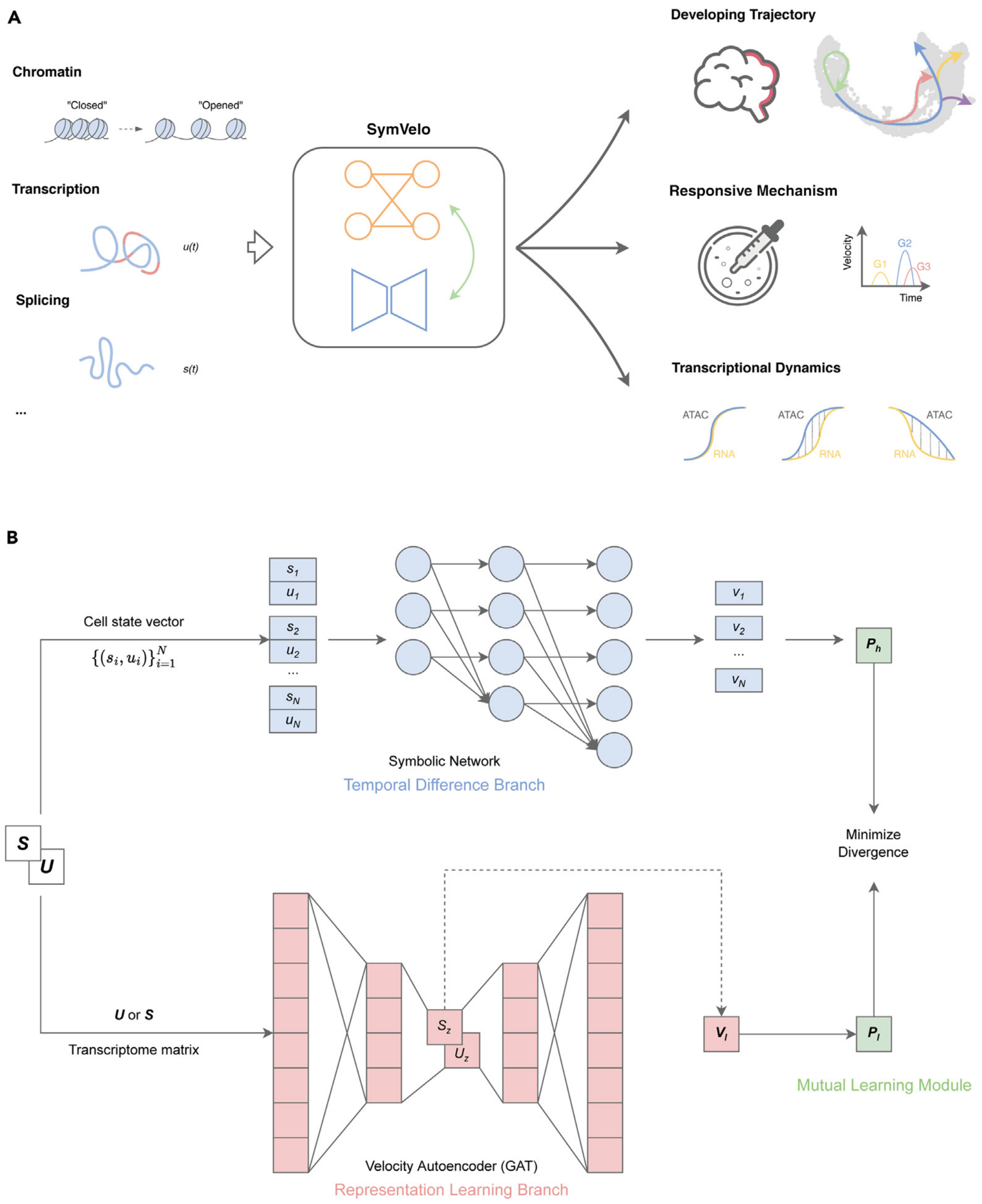


Figure 1. Profile of study and SymVelo's general framework

(A) SymVelo accepts multimodal data as input including scRNA-seq and scATAC-seq and applies the output velocity to several applications, such as trajectory inference.

(B) SymVelo consists of three modules: temporal difference module (blue), representation learning module (red), and mutual learning module (green).

learning module. The temporal difference module estimates high-dimensional RNA velocity in a continuous, bottom-up manner using a neural ODE. The core component, SymNet, is a customized symbolic network that employs a generalized kinetic model to represent transcriptional gene dynamics, and it can be optimized using data-driven techniques without imposing strict assumptions. The pre-trained representation learning module leverages an autoencoder to learn a low-dimensional representation of RNA velocity. To align the dimensionally mismatched velocities derived from the other two modules, the mutual learning module leverages Markov modeling of cell transitions.

To ascertain SymVelo's predictive accuracy in cell differentiation trajectories, this study contrasts its performance with that of scVelo (including stochastic and dynamical models) and VeloAE. This comparison is based on one simulated dataset and three real datasets. We utilize two metrics for performance assessment: the cross-boundary direction correctness score (CBDir) and in-cluster coherence (ICVCoh), as proposed by Chen Qiao et al.[7] Analysis of the tree-like simulated data reveals that SymVelo's velocity calculations exhibit remarkable time consistency, aligning closely with the "real" time clusters, as shown in Table S1 and Figure S2, thereby outperforming other methods.

Table 1. Quantitative benchmarking on three datasets with CBDir indices

Datasets	Direction	scVelo (stochastic)	scVelo (dynamic)	VeloAE	SymVelo (u/s)	SymVelo (c/u/s)
Dentate gyrus	nIPC, Neuroblast	0.814	0.906	0.916	0.099	–
	Neuroblast, Granule immature	0.627	−0.095	0.648	0.674	
	Granule immature, Granule mature	0.079	−0.157	−0.16	0.043	
	Radial Glia-like, Astrocytes	0.846	0.806	−0.743	0.449	
	OPC, OL	−0.886	−0.143	0.964	0.987	
	mean	0.296	0.264	0.325	0.45	
scNT-seq	0, 15	0.188	−0.011	0.57	0.761	
	15, 30	0.183	0.222	0.408	0.355	
	30, 60	0.261	0.276	0.283	0.338	
	60, 120	0.38	0.454	0.279	0.456	
	mean	0.253	0.235	0.385	0.477	
Multiome	Cyc., RG/Astro	0.13	0.264	0.272	−0.085	−0.155
	Cyc., nIPC/ExN	0.746	0.463	−0.391	0.527	0.623
	nIPC/ExN, ExM	0.146	0.102	−0.108	0.066	0.574
	ExM, ExUp	0.081	0.291	0.259	−0.282	0.075
	RG/Astro, mGPC/OPG	−0.232	0.037	0.371	0.486	0.259
	mean	0.174	0.231	0.081	0.142	0.275

Additionally, as evidenced in Tables 1 and S1 and Figure S1B, SymVelo consistently achieves the highest mean CBDir values across the three real datasets, while maintaining strong internal consistency within cell groups. These results highlight SymVelo's proficiency in mastering complex transcription dynamics and in providing an accurate depiction of continuous, high-dimensional RNA velocity and cell differentiation trajectories.

SymVelo delineates cell differentiation in the developing mouse dentate gyrus

Neurogenesis within the dentate gyrus endures throughout adulthood, during which progenitors or immature cells either proliferate or differentiate under regulation. To demonstrate the efficacy of SymVelo in revealing the direction of cell differentiation, we conduct experiments on the developing mouse dentate gyrus at two time points (postnatal day 12 and 35) and retain 2,930 cells and 2,000 highly variable genes (HVGs) for velocity estimation after preprocessing.[9] As demonstrated by the scores in Table 1, SymVelo accurately discerns the general orientation of cell differentiation relative to other methods. It successfully delineates the differentiation of the granule cell lineage, from neuroblast to mature granule cell, and captures the developmental process of isolated cell clusters, such as from oligodendrocyte precursor cell (OPC) to oligodendrocyte (OL).

We further elaborate on the characteristics of velocity inferred by SymVelo. As shown in Figure 2A, velocity can effectively distinguish distinct cell types, possibly because fully differentiated cell types, such as Cajal Retzius, or sub-lineage cell types, like OL, have more distinct genetic properties. The Eta squared coefficients of most genes in velocity exceed those in RNA (above the diagonal), indicating that velocity has a stronger correlation with cell type (Figure 2B, STAR Methods). Certain genes display higher or lower velocities in specific cell types in Figure 2A, which may be attributed to the skewed distribution of velocity. This assumption is confirmed, and we find that gene expressions are generally right skewed (skewness > 0, meaning a gene is highly expressed in a subset of cells), while the proportion of left-skewed and right-skewed distribution of velocity is comparable (Figure 2C). Since velocity can reflect changes in the abundance of unspliced and spliced RNA, the dynamic changes of transcripts can be further revealed in the time dimension by analyzing the correlation between velocity and RNA. The Spearman correlation test indicates that velocity and RNA of genes are positively correlated when the skewness is consistent and negatively correlated under inconsistent conditions. We analyze the pseudo-time trend of RNA and velocity of the top 10 and bottom 10 genes ordered by velocity skewness. Figure 2D demonstrates that RNA and velocity are generally consistent, with a few exceptions such as Gart and Ift140. Through the visualization of uniform manifold approximation and projection (UMAP) in Figure 2E, we show that several cells with high expression values are discrete, while a more continuous velocity is modeled. This suggests that velocity learned by SymVelo can mitigate the shortcomings of single-cell sequencing, such as snap-shot and drop-out phenomena.

The skewness in the data indicates the presence of differential genes. We conduct differential analysis on the velocity and RNA of genes, which led to the identification of four distinct sections. These sections are denoted as Velocity-Up-RNA-Up (UU), Velocity-Up-RNA-Down (UD), Velocity-Down-RNA-Up (DU), and Velocity-Down-RNA-Down (DD), as demonstrated in Figure 2F (STAR Methods). In instances where a gene's expression is down-regulated, the unspliced and spliced counts of cells are reduced, while they are increased when a gene is up-regulated. Moreover, cells

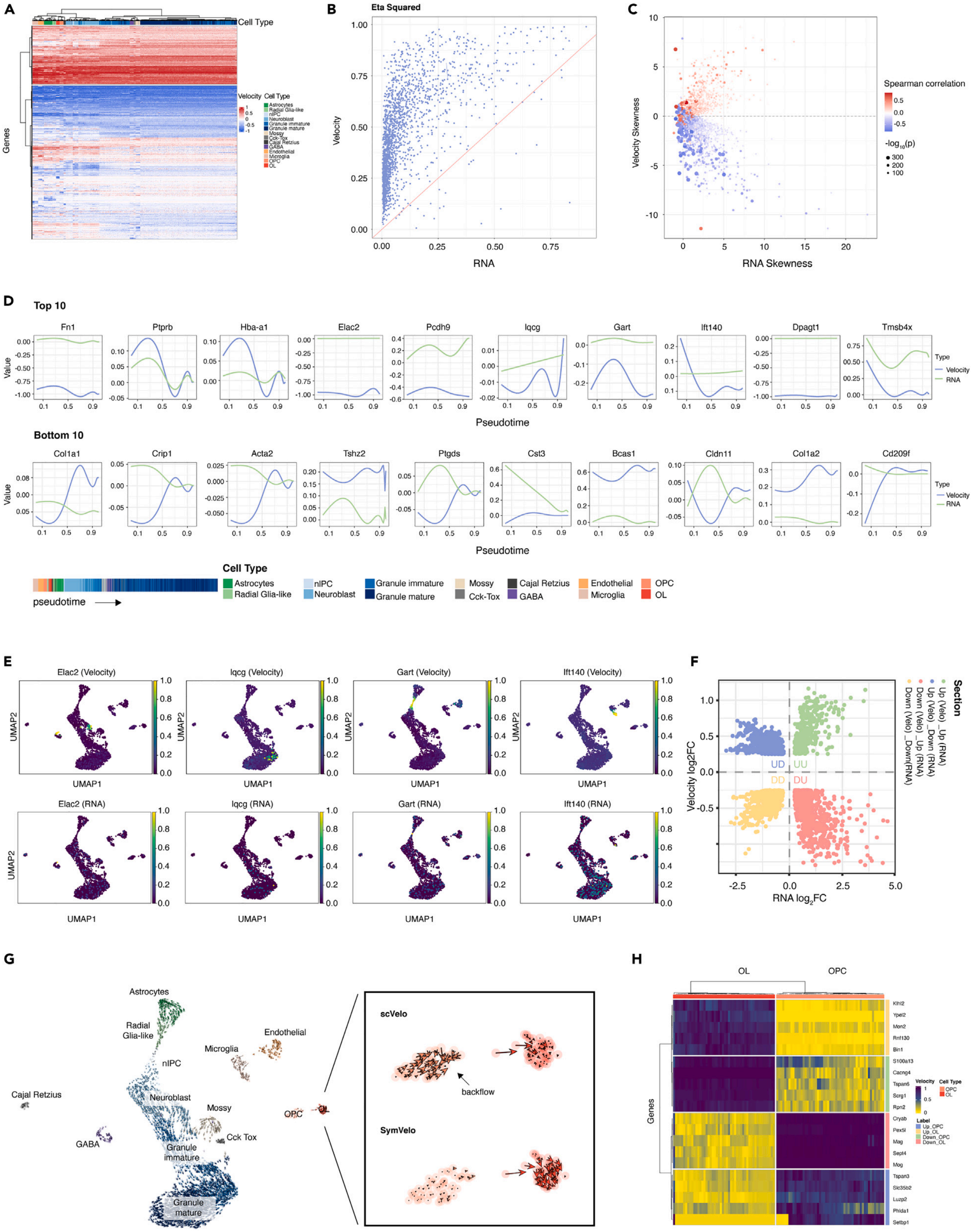

A
Cell Type
Genes
Velocity
1
0.5
0
-0.5
-1
Cell Type
Astrocytes
Radial Glia-like
nIPC
Neuroblast
Granule immature
Granule mature
Mossy
Cck-Tox
Cajal Retzius
GABA
Endothelial
Microglia
OPC
OL
B
Eta Squared
1.00
0.75
0.50
0.25
0.00
Velocity
0.00
0.25
0.50
0.75
RNA
C
10
5
0
-5
-10
Velocity Skewness
0
5
10
15
20
RNA Skewness
Spearman correlation
0.5
0.0
-0.5
-log10(p)
300
200
100
D
Top 10
Fn1
Ptprb
Hba-a1
Elac2
Pcdh9
Iqcg
Gart
Ift140
Dpagt1
Tmsb4x
Value
Type
Velocity
RNA
Pseudotime
Bottom 10
Col1a1
Crip1
Acta2
Tshz2
Ptgds
Cst3
Bcas1
Cldn11
Col1a2
Cd209f
Value
Type
Velocity
RNA
Pseudotime
Cell Type
Astrocytes
Radial Glia-like
nIPC
Neuroblast
Granule immature
Granule mature
Mossy
Cck-Tox
Cajal Retzius
GABA
Endothelial
Microglia
OPC
OL
pseudotime
E
Elac2 (Velocity)
Iqcg (Velocity)
Gart (Velocity)
Ift140 (Velocity)
Elac2 (RNA)
Iqcg (RNA)
Gart (RNA)
Ift140 (RNA)
UMAP2
UMAP1
1.0
0.8
0.6
0.4
0.2
0.0
F
Section
Up (Velo)_Up (RNA)
Up (Velo)_Down (RNA)
Down (Velo)_Up (RNA)
Down (Velo)_Down(RNA)
Velocity log2FC
1.0
0.5
0.0
-0.5
UD
UU
DD
DU
-2.5
0.0
2.5
5.0
RNA log2FC
G
Astrocytes
Radial Glia-like
nIPC
Endothelial
Microglia
Cajal Retzius
Neuroblast
Mossy
OPC
OL
GABA
Cck Tox
Granule immature
Granule mature
scVelo
backflow
SymVelo
H
OL
OPC
Genes
Klhl2
Ypel2
Mon2
Rnf130
Bin1
S100a13
Cacng4
Tspan5
Scrg1
Rpn2
Cryab
Pex5l
Mag
Sept4
Mog
Tspan3
Slc35b2
Luzp2
Phlda1
Setbp1
Velocity
1
0.5
0
Cell Type
OPC
OL
Label
Up_OPC
Up_OL
Down_OPC
Down_OL

Figure 2. Analysis of the dentate gyrus Dataset

(A) Cell type clustering based on velocity. Each column represents a cell, which each row represents a gene. We color the cells by their cell type at the top annotation of the heatmap. The normalized velocity is used to plot the main body of the heatmap.

(B) Calculation of Eta squared between cell type and either RNA (x axis) or velocity (y axis). Each point in the plot represents a gene, and the red line is the equation y = x.

(C) Comparison of skew distributions of velocity or RNA values. Each point in the plot represents a gene and is colored by the Spearman correlation coefficient between velocity and RNA values. The size of points is inversely proportional to log10(p) value.

(D) Changes in velocity (blue line) or RNA (green line) along pseudo-time. The top 10 and bottom 10 genes, ordered by velocity skewness, are shown. Pseudo-time calculated by velocity is shown as the x axis. The distribution of cell types over pseudo-time is visualized at the bottom of the plot.

(E) Inconsistencies between RNA and velocity of several genes are visualized by UMAP colored by normalized velocity or expression values.

(F) Four sections based on differential genes in certain cell types of velocity and RNA. The x axis on the figure represents the log 2 logarithmic transformation of fold change between cell types in expression level while y axis is in velocity values. The color of points is generated from the fold change value of RNA expression and velocity.

(G) Streamline plot (left) suggests the developmental processes of dentate gyrus calculated on velocity. In the right box of plot, differentiation flow of oligodendrocyte lineage is zoomed in for comparison of scVelo and SymVelo, and the backflow of scVelo is highlighted.

(H) Differential genes in oligodendrocyte precursor cells (OPCs) or myelinating oligodendrocytes (OLs). The genes selected by differential expression analysis (one versus others) are annotated as up-regulated or down-regulated in the right of the panel. The gradient of colors reflects the normalized velocity.

that exhibit an increased velocity are more likely to be above the steady state, as shown in Figure S3A. Gene ontology analysis of each section in Figure S3B reveals that genes in the UD section are related to neurotransmission, such as synapse signaling, while genes in the DU section are associated with cell proliferation and differentiation. The genes within the DU section, including Gad2, Sept4, and Hmgb1, display significant peaks and valleys (in GABAergic neurons, OL, and neurogenic intermediate progenitor cell, respectively), as demonstrated in Figures S4A and S4B. Hmgb1 is a crucial factor in neurogenesis and is known to affect the proliferation and differentiation of stem cells and progenitor cells.[10] Gad2 controls the synthesis of GABA,[11],[12] and Sept4 constitutes the primary cytoskeleton component of mature myelin, thus impacting myelination in OL.[13] Although the velocity in the DU section decreases, the expression level remains relatively high. We further analyze specific differential genes of velocity, as presented in Figure S4C, including Gart, Ret, and Olfm2, which mediate neural stem cell and progenitor cell proliferation and expansion,[14],[15] nucleosome acetylation,[16] and maturation of OLs,[17] respectively.

Notably, our proposed approach, SymVelo, learns velocity with more consistent direction and without backflow in OL lineage, as demonstrated in Figures 2G and Table 1. Differential genes in Figure 2H can efficiently distinguish between two cell types. For instance, Klhl2 is constitutively expressed in developing and mature OLs and is up-regulated during OL differentiation.[18] Mon2 and Dopey1 jointly mediate OL myelination,[19] while in OPC, Tspan3 is up-regulated, promoting cell proliferation. As cell differentiation proceeds, the velocity of Tspan3 decelerates and facilitates the migration of OL.[20],[21] In contrast, CACNG4 is associated with TNFRSF21, which controls the maturation of immature OL.[22] Mag and Mog are important components of the membrane surface of OL.[23]

Through a continuous vector field, SymVelo exhibits robustness to data sparsity, providing more power in the discovery of cell identity or differential genes.

SymVelo elucidates neurons response mechanism under stimulation

Recent study has highlighted the role of neuronal activity in inducing cell type-specific genetic changes. Continuous stimulation has been shown to trigger time-dependent gene responses, which can be effectively captured by RNA velocity.[24] However, Qi Qiu et al.[24] have noted an inconsistency between the RNA velocity flow and the directionality, dependent on neuronal activity. To address this, we investigate whether SymVelo's splicing kinetics-based RNA velocity can infer the transcriptional state trajectories of single cells in response to neuronal activation, using a scNT-seq dataset of 3,066 high-quality excitatory neurons stimulated by potassium chloride (KCl) for various durations of neuronal activity (0-, 15-, 30-, 60-, and 120-min). Our analysis confirms the lack of consistency between observed velocity direction and stimulation time in different modes using scVelo, as previously observed by Qi Qiu et al., while SymVelo's velocity flow aligns with the stimulation time interval (Figures 3A and S5A). We also conduct comprehensive ablation studies to justify SymVelo's superior performance compared to scVelo and a random module, attributing this to SymVelo's mutual module which processes the cell pair of the current and future cell states derived from pre-trained representation learning module (Figure S5C). The Eta squared coefficient confirms a higher correlation between SymVelo's velocity and time (Figures 3B and S5B).

As the duration of stimulation time increases, activity-regulated genes exhibit varying response patterns. In a previous study, Qi Qiu et al. selected 24 early-response genes and 73 late-response genes that are consistent with the duration of stimulus, as shown in Figure 3C.[24] To further illustrate the performance of velocity at the gene level, we calculate the Spearman correlation between the velocity of each gene and the early- or late-response gene expression level. Figure 3D shows that the velocity of 62 response genes (only HVGs) determined by SymVelo achieves higher correlation, while the velocity from scVelo is less correlated. We apply a threshold of 0.4 for filtering. Moreover, we find that the genes identified by SymVelo have a higher overlap with those reported in the article (Figure 3E; Figure S5D). Notably, SymVelo also identifies novel response genes (Figure 3F, top: spearman coefficient, bottom: expression of some genes). For example, Ppp1r10 can strongly bind to protein phosphatase 1 (PP-1) to inhibit PP-1-mediated dephosphorylation of substrates and regulate synaptic transmission and plasticity. Prolonged depolarization leads to changes in calcium ion channels and the activation of some kinases such as calcium–calmodulin (CaM)-dependent protein kinase II (CaMKII), which can inhibit PP-1 through downstream substrates.[25–27] As the influx of calcium ions changes the stability of neuronal actin, Ivns1abp can stabilize the actin skeleton and protect neurons from cell death induced by actin instability.[28],[29]

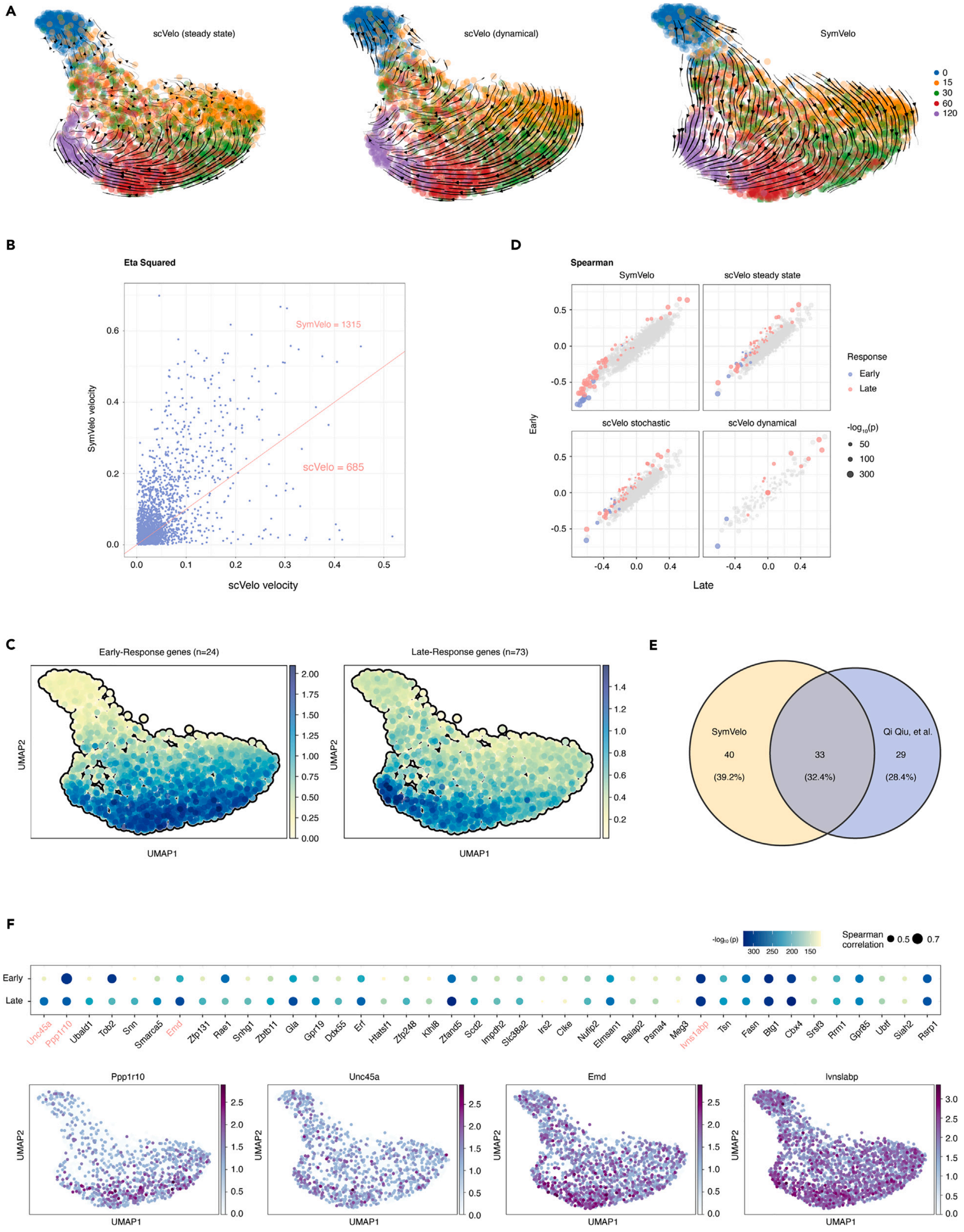

A
scVelo (steady state)
scVelo (dynamical)
SymVelo
0
15
30
60
120
B
Eta Squared
SymVelo = 1315
scVelo = 685
SymVelo velocity
scVelo velocity
0.0
0.2
0.4
0.6
0.0
0.1
0.2
0.3
0.4
0.5
D
Spearman
SymVelo
scVelo steady state
scVelo stochastic
scVelo dynamical
Early
Late
0.5
0.0
-0.5
-0.4
0.0
0.4
Response
Early
Late
-log10(p)
50
100
300
C
Early-Response genes (n=24)
Late-Response genes (n=73)
UMAP2
UMAP1
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
1.4
1.2
1.0
0.8
0.6
0.4
0.2
E
SymVelo
40
(39.2%)
33
(32.4%)
Qi Qiu, et al.
29
(28.4%)
F
-log10 (p)
300 250 200 150
Spearman correlation
0.5
0.7
Early
Late
Unc45a
Ppp1r10
Ubald1
Tob2
Snn
Smarca5
Emd
Zfp131
Rae1
Snhg1
Zbtb11
Gla
Gpr19
Ddx55
Erf
Htatsf1
Zfp248
Klhl8
Zfand5
Scd2
Impdh2
Slc38a2
Irs2
Clke
Nufip2
Elmsan1
Baiap2
Psma4
Meg3
Ivns1abp
Tsn
Fasn
Btg1
Cbx4
Srsf3
Rrm1
Gpr85
Ubtf
Siah2
Rsrp1
Ppp1r10
Unc45a
Emd
Ivnslabp
UMAP2
UMAP1
2.5
2.0
1.5
1.0
0.5
0.0
3.0

Figure 3. Identification of response genes in neuronal activity by SymVelo

(A) Streamline plot of three methods, including scVelo steady-state mode, scVelo dynamical mode, and SymVelo. Points are colored by the duration of neuronal activity.

(B) Eta squared measures the correlation between velocity and cell type. Values of x axis are the Eta squared between scVelo and cell type, while values of y axis are the Eta squared between SymVelo and cell type. The text colored red in the plot indicates the total number of genes in which one algorithm correlates more than the other algorithm, and the red line represents the equation y = x.

(C) UMAP of 3,066 high-quality excitatory neurons colored by mean new expression levels of early- and late-response genes among cells.

(D) Spearman correlation between velocity and average expression levels of early- (y axis) and late-response (x axis) genes. Only HVGs are selected leading to 62 overlapping response genes. Each point represents a gene which is colored according to their response time, including blue and red for early- and late-response genes, respectively. The size of points is proportional to –log10(p value).

(E) Venn plot of consistency of gene sets detected by SymVelo using 0.4 as cutoff of Spearman coefficient and reported in the previous study.

(F) 40 novel genes identified by SymVelo are listed in detail. At the top panel, the size of points is proportional to spearman correlation between gene's velocity and expression of early- or late-response genes, and the color is related to the p value. Several genes are highlighted with red font color as examples and visualized as UMAP based on RNA expression values at the bottom panel.

Furthermore, UNC-45A, a chaperone of conventional and unconventional myosin, mediates contractile force and actin-based motility critical for proper growth cone motility and neurite extension.[30] Lastly, EMD-encoded Emerin is involved in nuclear calcium homeostasis.[31]

Overall, despite many activity-mediated genes having rapid splicing kinetics, SymVelo can efficiently restore and reveal neuronal activity-dependent gene response mechanisms.

SymVelo discovers distinctive transcription pattern based on multi-modality

Advancements in technology have made it possible to measure multiple modalities simultaneously in a single cell. To utilize the information from different modalities, SymVelo is designed with a more flexible structure. We focus on a human cerebral cortex dataset, employing the 10× Multiome to simultaneously profile gene expression and chromatin status in the same cell, to examine its performance.[32] The result in Figure 4A illustrates that, if only a single modality (i.e., scRNA-seq) is considered, the velocity flow appears to be countercurrent, such as transitioning from maturing neurons (ExM) to intermediate progenitor or newborn excitatory neurons (nIPC/ExN). This discrepancy could be attributed to the high complexity and high sparseness of multimodal sequencing technology. However, by integrating chromatin accessibility data (scATAC-seq), the velocity flow aligns with expectations and the predicted pseudo-time is more consistent with the maturation direction of neurons.

To understand the impact of chromatin accessibility on velocity, we analyze the trend of chromatin accessibility and splicing over pseudo-time. Considering the interference of different sub-lineages, such as astrocyte and OL, in time-dependent analysis and the unclear flow directions of early-formed layers such as subplate and deeper layer, we select a subset of cells, including cycling progenitors (Cyc.), nIPC, ExM, and upper-layer neurons (ExUp) for subsequent analysis. Soft clustering is performed to identify patterns of chromatin and transcription trends for each gene (Figure S6 and STAR Methods). We then combine three gene features (i.e., spliced, unspliced, and chromatin accessible values) to investigate the transcriptional paradigms. Our statistical analysis reveals several prominent patterns (Figure 4B), namely c6_u1_s6, c5_u5_s4, c3_u2_s2, and c1_u1_s6 (c, u, and s represent chromatin, unspliced, and spliced, respectively). As illustrated In Figures 4C and S7, genes in the c6_u1_s6 group, such as TOP2A and CDH13, exhibit a higher coupling between the chromatin and the transcriptional state, suggesting that the transcriptional initiation stage of these genes is rapid or requires fewer regulatory elements. Furthermore, we observe a lag between two states of some genes in the other three groups. For instance, the chromatin of genes in the c5_u5_s4 group, such as ARFGEF3 and ATAD2, remains open when the transcription is down-regulated. Similarly, the accessibility of genes in the c1_u1_s6 group, like CSMD1, is open for a period of time in advance. These findings suggest that the transcription efficiency of genes is heterogeneous and that chromatin accessibility, transcription initiation, and splicing contribute to this heterogeneity, consistent with previous assumptions.[33]

The SymVelo model is not restricted to single-modal data but can also accommodate two or more modalities with ease. The inclusion of multiple layers of information sheds new light on cell differentiation, facilitating a more comprehensive understanding of cellular processes.

DISCUSSION

Our objective is to introduce a unified framework for estimating RNA velocity that amalgamates the strengths of low- and high-dimensional representations from both discrete and continuous dynamic system perspectives. To achieve this goal, we propose SymVelo, which constructs the continuous vector field through a symbolic neural network in the temporal difference branch. This approach distinguishes itself from methodologies like Dynamo,[34] which reconstructs the vector field after estimating the velocity. We posit that our method offers more precise prior information, thereby optimizing the framework more effectively. Moreover, the inclusion of SymNet in our model enables comparisons with other models, such as DeepVelo[6] and Dynamo.[34]

SymVelo excels at illustrating multi-lineage systems by extending the original first-order equation into high-order polynomials. In the conventional implementation,[1] constant kinetic parameters are assumed for a single gene across different cells. However, this simplification has been recently expanded by Haotian Cui et al.,[35] who introduced neural networks to approximate the assessment of kinetic parameters. Our proposed framework adopts a similar strategy, estimating high-order kinetic parameters in a manner that is aligned with the design of SymNet, considering the entirety of the expression state.

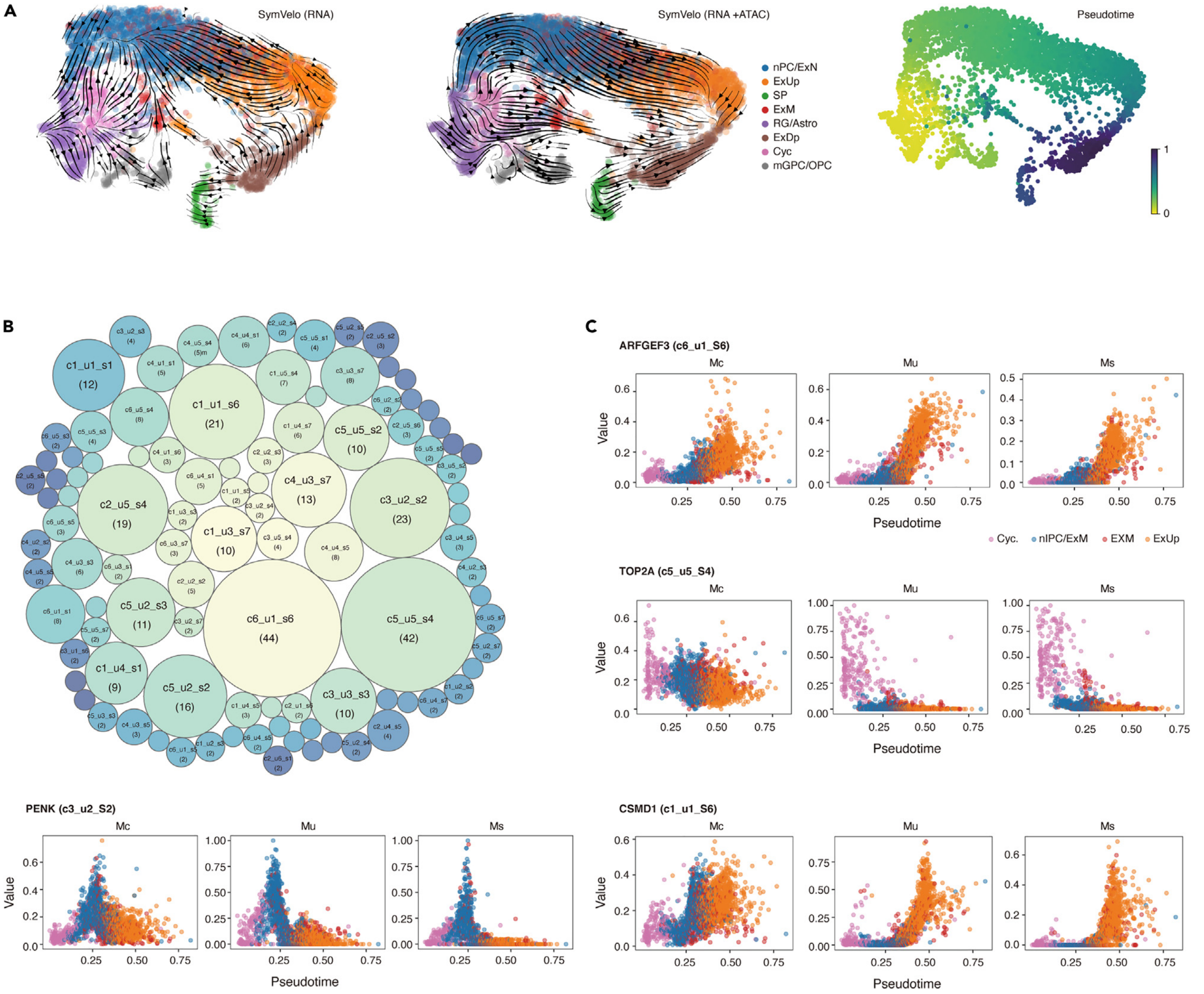


Figure 4. Transcriptional patterns of the human cortex multimodal dataset

(A) Velocity streamline plots colored by cell types with (left) and without (middle) chromatin accessibility data by SymVelo. Pseudo-time of each cell calculated from velocity is visualized in the UMAP (right). nIPC/ExN: intermediate progenitor cells/newborn excitatory neurons; ExUp: upper-layer neurons; SP: subplate; ExM: maturing neurons; RG/Astro: radial glia/astrocytes; ExDp: deep-layer neurons; Cyc.: cycling progenitors; mGPC/OPC: multipotent glial progenitor cells/ oligodendrocyte progenitor cells.

(B) Statistics of all combinatorial patterns of chromatin and transcriptional trends. The individual trends are depicted in Figure S5. The numbers in parentheses indicate the number of genes that belong to the group. c, u, and s represent chromatin, unspliced, and spliced features, respectively, followed by the cluster numbers of each trend. For example, "c6_u1_s6" means combination of cluster 6 of chromatin feature, cluster 1 of unspliced feature, and cluster 6 of spliced feature. The size and color are proportional to the number.

(C) Four prominent patterns are illustrated with examples plotted as time-dependent trends of chromatin, unspliced, and spliced values. X axis is the pseudo-time calculated from SymVelo's velocity, and y axis is the first-order moments of spliced, unspliced, and chromatin feature. Each point represents a cell colored by its cell type.

In the context of multi-omics scenarios, our framework presents a flexible solution. The input to SymNet allows for abstraction of the number of modalities to a flexible dimension, which facilitates easy extension to multi-omics scenarios with minimal modifications. Additionally, SymNet's transparency offers a window to infer the analytic form of the dynamics underlying the system. This level of abstraction and adaptability represents a significant stride in RNA velocity estimation, with the potential for advancements in multi-omics research and beyond.

Furthermore, the field is witnessing the emergence of various RNA velocity calculation methods based on artificial intelligence, such as Velo-Predictor[36] and LatentVelo.[8] This accentuates the necessity of constructing a more comprehensive and rationale comparative framework when developing state-of-the-art RNA velocity techniques, which encompass an evaluation of the model type. Despite the presence of quantitative indicators like CBDir that address the shortcomings of previous metrics, they too have limitations, including susceptibility to cell annotation and dimensionality reduction methods. Hence, there is an urgent need to develop more reliable indicators for evaluating different methods.

In conclusion, our proposed SymVelo exhibits notable adaptability and interpretability in various scenarios. The inclusion of symbolic neural network provides precision and transparency, elucidating the dynamics of underlying systems. Its versatile design allows for easy adaptability to multi-omics scenarios, broadening its applicability. Despite its adaptability, SymVelo retains biological interpretability, offering accurate depictions of multi-lineage systems. We believe that SymVelo holds significant potential to drive future advancements in RNA velocity estimation and related single-cell transcriptomics research.

Limitations of the study

This study confronts several potential limitations. Firstly, the dataset employed, particularly the multimodal dataset, is inadequate. This insufficiency impedes the effective demonstration of SymVelo's performance. To address this shortfall, we aim to enrich the dataset in future research endeavors, enabling a more comprehensive assessment of the model's capabilities. Secondly, our use of CBDir for quantitatively assessing RNA velocity raises concerns regarding its sensitivity to cell annotation and dimensionality reduction methods. Relying solely on CBDir for performance evaluation is thus considered inadequate, necessitating the development of more robust evaluation criteria in future studies. Furthermore, SymVelo's dependence on a modified VeloAE providing the future state of cells introduces a susceptibility to the limitations inherent in VeloAE. With the growing availability of deep learning-based methods for RNA velocity calculation, exploring alternative approaches to provide future states may yield more dependable outcomes.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.109635.

ACKNOWLEDGMENTS

This research is supported by Guangdong Life Science Instruments and Equipment Engineering Technology Research Center, MGI Tech, Shenzhen 518083, China.

AUTHOR CONTRIBUTIONS

M. Yang conceived the problem and designed the study. Y.Y. and Q.H. performed bioinformatics analysis. M. Yuan, H.Y., and C.X. performed algorithm design and deep learning experiments. M. Yuan, H.Y., Y.Y., Q.H., and C.X. wrote the manuscript. M. Yang and L.Z. jointly supervised the work.

DECLARATION OF INTERESTS

M. Yang is an employee and shareholder of MGI, BGI-Shenzhen.

Received: August 18, 2023
Revised: December 4, 2023
Accepted: March 26, 2024
Published: March 28, 2024

REFERENCES

1. Bergen, V., Lange, M., Peidli, S., Wolf, F.A., and Theis, F.J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. Nat. Biotechnol. *38*, 1408–1414. https://doi.org/10.1038/s41587-020-0591-3.
2. La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature *560*, 494–498. https://doi.org/10.1038/s41586-018-0414-6.
3. Bergen, V., Soldatov, R.A., Kharchenko, P.V., and Theis, F.J. (2021). RNA velocity—current challenges and future perspectives. Mol. Syst. Biol. *17*, e10282. https://doi.org/10.15252/msb.202110282.
4. Gu, Y., Blaauw, D., and Welch, J.D. (2022). Bayesian inference of rna velocity from multi-lineage single-cell data. Preprint at bioRxiv. https://doi.org/10.1101/2022.07.08.499381.
5. Gayoso, A., Weiler, P., Lotfollahi, M., Klein, D., Hong, J., Streets, A., Theis, F.J., and Yosef, N. (2024). Deep generative modeling of transcriptional dynamics for RNA velocity analysis in single cells. Preprint at bioRxiv. https://doi.org/10.1038/s41592-023-01994-w.
6. Chen, Z., King, W.C., Hwang, A., Gerstein, M., and Zhang, J. (2022). DeepVelo: Single-cell transcriptomic deep velocity field learning with neural ordinary differential equations. Sci. Adv. *8*, eabq3745. https://doi.org/10.1126/sciadv.abq3745.
7. Qiao, C., and Huang, Y. (2021). Representation learning of RNA velocity reveals robust cell transitions. Proc. Natl. Acad. Sci. USA *118*. e2105859118. https://doi.org/10.1073/pnas.2105859118.
8. Farrell, S., Mani, M., and Goyal, S. (2023). Inferring single-cell transcriptomic dynamics with structured latent gene expression dynamics. Cell Reports Methods *3*. https://doi.org/10.1016/j.crmeth.2023.100581.
9. Hochgerner, H., Zeisel, A., Lönnerberg, P., and Linnarsson, S. (2018). Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell RNA sequencing. Nat. Neurosci. *21*, 290–299. https://doi.org/10.1038/s41593-017-0056-2.
10. Zhao, X., Rouhiainen, A., Li, Z., Guo, S., and Rauvala, H. (2020). Regulation of neurogenesis in mouse brain by HMGB1. Cells *9*, 1714. https://doi.org/10.3390/cells9071714.
11. Pan, Z.Z. (2012). Transcriptional control of Gad2. Transcription *3*, 68–72. https://doi.org/10.4161/trns.19511.
12. Lee, C.L., Lam, K.K.W., Vijayan, M., Koistinen, H., Seppala, M., Ng, E.H.Y., Yeung, W.S.B., and Chiu, P.C.N. (2016). The pleiotropic effect of glycodelin-A in early pregnancy. American Journal of Reproductive Immunology *75*, 290–297. https://doi.org/10.1111/aji.12471.
13. Patzig, J., Dworschak, M.S., Martens, A.-K., and Werner, H.B. (2014). Septins in the glial cells of the nervous system. Biol. Chem. *395*, 143–149. https://doi.org/10.1515/hsz-2013-0240.
14. Yang, Y., Song, L., Huang, X., Feng, Y., Zhang, Y., Liu, Y., Li, S., Zhan, Z., Zheng, L., Feng, H., and Li, Y. (2021). PRPS1-mediated purine biosynthesis is critical for pluripotent stem cell survival and stemness. Aging (Albany NY) *13*, 4063–4078. https://doi.org/10.18632/aging.202372.
15. Vickers, N.J. (2017). Animal communication: when i'm calling you, will you answer too? Curr. Biol. *27*, R713–R715. https://doi.org/10.1016/j.cub.2017.05.064.
16. Harding, E.C., Franks, N.P., and Wisden, W. (2019). The temperature dependence of sleep. Front. Neurosci. *13*, 336. https://doi.org/10.3389/fnins.2019.00336.
17. Sultana, A., Nakaya, N., Senatorov, V.V., and Tomarev, S.I. (2011). Olfactomedin 2: Expression in the Eye and Interaction with Other Olfactomedin Domain–Containing Proteins. Investigative ophthalmology & visual science *52*, 2584–2592. https://doi.org/10.1167/iovs.10-6356.
18. Dhanoa, B.S., Cogliati, T., Satish, A.G., Bruford, E.A., and Friedman, J.S. (2013). Update on the Kelch-like (KLHL) gene family. Hum. Genom. *7*, 1–7. https://doi.org/10.1186/1479-7364-7-13.
19. Gong, C., Kim, E.M., Wang, Y., Lee, G., and Zhang, X. (2019). Multiferroicity in atomic van der Waals heterostructures. Nat. Commun. *10*, 2657. https://doi.org/10.1038/s41467-019-10693-0.
20. Yang, Y.-G., Sari, I.N., Zia, M.F., Lee, S.R., Song, S.J., and Kwon, H.Y. (2016). Tetraspanins: Spanning from solid tumors to hematologic malignancies. Experimental hematology *44*, 322–328. https://doi.org/10.1016/j.exphem.2016.02.006.
21. Tiwari-Woodruff, S.K., Buznikov, A.G., Vu, T.Q., Micevych, P.E., Chen, K., Kornblum, H.I., and Bronstein, J.M. (2001). OSP/claudin-11 forms a complex with a novel member of the tetraspanin super family and β1 integrin and regulates proliferation and migration of oligodendrocytes. The J. of cell Biol. *153*, 295–306. https://doi.org/10.1083/jcb.153.2.295.
22. Göteson, A., Isgren, A., Jonsson, L., Sparding, T., Smedler, E., Pelanis, A., Zetterberg, H., Jakobsson, J., Pålsson, E., Holmén-Larsson, J., and Landén, M. (2021). Cerebrospinal fluid proteomics targeted for central nervous system processes in bipolar disorder. Mol. Psychiatry *26*, 7446–7453. https://doi.org/10.1038/s41380-021-01236-5.
23. Ambrosius, W., Michalak, S., Kozubski, W., and Kalinowska, A. (2020). Myelin oligodendrocyte glycoprotein antibody-associated disease: current insights into the disease pathophysiology, diagnosis and management. Int. J. Mol. Sci. *22*, 100. https://doi.org/10.3390/ijms22010100.
24. Qiu, Q., Hu, P., Qiu, X., Govek, K.W., Cámara, P.G., and Wu, H. (2020). Massively parallel and time-resolved RNA sequencing in single cells with scNT-seq. Nat. Methods *17*, 991–1001. https://doi.org/10.1038/s41592-020-0935-4.
25. Platholi, J., and Hemmings, H.C., Jr. (2021). Modulation of dendritic spines by protein phosphatase-1. Adv. Pharmacol. *90*, 117–144. https://doi.org/10.1016/bs.apha.2020.10.001.
26. Choy, M.S., Hieke, M., Kumar, G.S., Lewis, G.R., Gonzalez-DeWhitt, K.R., Kessler, R.P., Stein, B.J., Hessenberger, M., Nairn, A.C., Peti, W., and Page, R. (2014). Understanding the antagonism of retinoblastoma protein dephosphorylation by PNUTS provides insights into the PP1 regulatory code. Proc. Natl. Acad. Sci. USA *111*, 4097–4102. https://doi.org/10.1073/pnas.1317395111.
27. Seth, A.K., Barrett, A.B., and Barnett, L. (2015). Granger causality analysis in neuroscience and neuroimaging. J. Neurosci. *35*, 3293–3297. https://doi.org/10.1523/JNEUROSCI.4399-14.2015.
28. Fujii, R., and Takumi, T. (2005). TLS facilitates transport of mRNA encoding an actin-stabilizing protein to dendritic spines. J. Cell Sci. *118*, 5755–5765. https://doi.org/10.1242/jcs.02692.
29. Cristofanilli, M., and Akopian, A. (2006). Calcium channel and glutamate receptor activities regulate actin organization in salamander retinal neurons. J. Physiol. *575*, 543–554. https://doi.org/10.1113/jphysiol.2006.114108.
30. Kanfer, G., Peterka, M., Arzhanik, V.K., Drobyshev, A.L., Ataullakhanov, F.I., Volkov, V.A., and Kornmann, B. (2017). CENP-F couples cargo to growing and shortening microtubule ends. Mol. Biol. Cell *28*, 2400–2409. https://doi.org/10.1091/mbc.e16-11-0756.
31. Shimojima, M., Yuasa, S., Motoda, C., Yozu, G., Nagai, T., Ito, S., Lachmann, M., Kashimura, S., Takei, M., Kusumoto, D., et al. (2017). Emerin plays a crucial role in nuclear invagination and in the nuclear calcium transient. Sci. Rep. *7*, 44312. https://doi.org/10.1038/srep44312.
32. Trevino, A.E., Müller, F., Andersen, J., Sundaram, L., Kathiria, A., Shcherbina, A., Farh, K., Chang, H.Y., Paşca, A.M., Kundaje, A., et al. (2021). Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. Cell *184*, 5053–5069.e23. https://doi.org/10.1016/j.cell.2021.07.039.
33. Li, C., Virgilio, M.C., Collins, K.L., and Welch, J.D. (2023). Multi-omic single-cell velocity models epigenome–transcriptome interactions and improves cell fate

prediction. Nat. Biotechnol. *41*, 387–398. https://doi.org/10.1038/s41587-022-01476-y.

34. Qiu, X., Zhang, Y., Martin-Rufino, J.D., Weng, C., Hosseinzadeh, S., Yang, D., Pogson, A.N., Hein, M.Y., Hoi Joseph Min, K., Wang, L., et al. (2022). Mapping transcriptomic vector fields of single cells. Cell *185*, 690–711.e45. https://doi.org/10.1016/j.cell.2021.12.045.
35. Cui, H., Maan, H., Vladoiu, M.C., Zhang, J., Taylor, M.D., and Wang, B. (2024). DeepVelo: deep learning extends RNA velocity to multi-lineage systems with cell-specific kinetics. Genome Biol. *25*, 27. https://doi.org/10.1186/s13059-023-03148-9.
36. Wang, X., and Zheng, J. (2021). Velo-Predictor: an ensemble learning pipeline for RNA velocity prediction. BMC Bioinf. *22*, 1–14. https://doi.org/10.1186/s12859-021-04330-1.
37. Zhang, Z., Wu, R., Gui, M., Jiang, Z., and Li, P. (2021). VeloSim: Simulating single cell gene-expression and RNA velocity. Preprint at bioRxiv. https://doi.org/10.1101/2021.01.11.426277.
38. Lange, M., Bergen, V., Klein, M., Setty, M., Reuter, B., Bakhti, M., Lickert, H., Ansari, M., Schniering, J., Schiller, H.B., et al. (2022). CellRank for directed single-cell fate mapping. Nat. Methods *19*, 159–170. https://doi.org/10.1038/s41592-021-01346-6.
39. Li, T., Shi, J., Wu, Y., and Zhou, P. (2020). On the mathematics of RNA Velocity I: theoretical analysis. Preprint at bioRxiv. https://doi.org/10.1101/2020.09.19.304584.
40. Long, Z., Lu, Y., and Dong, B. (2019). PDE-Net 2.0: Learning PDEs from data with a numeric-symbolic hybrid deep network. J. Comput. Phys. *399*, 108925. https://doi.org/10.1016/j.jcp.2019.108925.
41. Kipf, T.N., and Welling, M. (2016). Semi-supervised Classification with Graph Convolutional Networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1609.02907.
42. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). Graph Attention Networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1710.10903.
43. Zhang, Y., Xiang, T., Hospedales, T.M., and Lu, H. (2018). Deep Mutual Learning. Preprint at arXiv. 4320–4328. https://arxiv.org/abs/1706.00384.
44. Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the Knowledge in a Neural Network. Preprint at arXiv. https://doi.org/10.48550/arXiv.1503.02531.
45. Kumar, L., and E Futschik, M. (2007). Mfuzz: a software package for soft clustering of microarray data. Bioinformation *2*, 5–7. https://doi.org/10.6026/97320630002005.
46. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., III, Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573–3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.
47. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation *2*, 100141. https://doi.org/10.1016/j.xinn.2021.100141.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Code and data for development and evaluation	This paper	https://github.com/melobio/SymVelo
Software and algorithms		
scVelo (v0.2.5)	Bergen et al.[1]	https://github.com/theislab/scvelo
VeloAE (v0.2.0)	Chen et al.[7]	https://github.com/qiaochen/VeloAE
Python (version 3.8.5)	Python Software Foundation	https://www.python.org/
PyTorch (version 1.9.0)	Python package	https://pytorch.org/
R (version 4.1.2)	R software	http://www.R-project.org
Seurat (version 4.1.1)	R package	https://github.com/satijalab/seurat
Mfuzz (version 2.54.0)	R package	http://mfuzz.sysbiolab.eu/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Meng Yang (yangmeng1@mgi-tech.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Three real datasets leveraged in this study have been deposited at GitHub (https://github.com/melobio/SymVelo) and are publicly available as of the date of publication. Specifically, the dentate gyrus dataset is obtained using the "scv.datasets.dentategyrus" function from scVelo, while the scNT-seq dataset is sourced upon request from Qi Qiu.[24] Additionally, lists of early- and late-response genes of scNT-seq dataset are derived from Table S2 in the referenced publication.[24] Furthermore, the Multiome dataset originates from MultiVelo.[33] For simulated data, we adhere to the protocol outlined in VeloSim's tutorial "Simulation for tree-like trajectory".[37]
- SymVelo is implemented in Python. The package has been deposited at GitHub (https://github.com/melobio/SymVelo) and is publicly available as of the date of publication.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Input data & preprocessing

The data used in this study were either obtained from the primary research article or requested directly from the authors. To analyze the scRNA-seq data, SymVelo requires input of both spliced and unspliced counts. Initially, genes with less than 30 counts (both unspliced and spliced) were excluded, followed by a logarithmic transformation. For further analysis, we select 2000 highly variable genes (HVGs). The first order moments of unspliced and spliced counts for each cell, in relation to its nearest neighbors, were calculated using scVelo. This step is essential for deterministic velocity estimation.

Development of a generalized kinetic model for gene dynamics

In previous frameworks for RNA velocity quantification,[1,2] the gene-specific transcription dynamics are formulated by a linear first-order autonomous ODE with constant rate parameters,

$$\frac{du}{dt} = \alpha(t) - \beta u(t) \quad \text{(Equation 1)}$$

$$\frac{ds}{dt} = \beta u(t) - \gamma s(t) \quad \text{(Equation 2)}$$

where u(t), s(t) represents normalized unspliced and spliced mRNA reads, and the reaction dynamics for each gene are depicted temporally by transcription rate α(t), splicing rate β, and degradation rate γ. RNA velocity[2] is then defined as the derivative of spliced abundance s(t), which recovers directed information from transcriptional dynamics. Velocities across genes could be combined to extrapolate the future state of an individual cell.

However, due to complex biological mechanisms that modulate transcription, splicing, and degradation rates, the above first-order equation has been demonstrated to be inadequate to describe realistic transcriptional dynamics. Additionally, the gene-specific model omits the interaction among different genes.

To address these limitations, we propose a generalized kinetic model that discards the first-order assumption and gene-specific setting, while only retaining the reasonable autonomous property. Formally, suppose a vector $x \in R^{M \cdot d}$ denotes a cell's expression state, where M is the number of genes and d is the dimension of the gene-specific cell state. In this scenario, d = 2, which means the cell state can be viewed as a concatenation of unspliced and spliced counts $u,s \in R^M$, i.e., $x = (u,s)$. We aim to model the transcriptional kinetics with the following autonomous dynamic system,

$$\dot{x} = F(x) \tag{Equation 3}$$

where F(·) describes the continuous vector field.[34] To model complex transcriptional scenarios,[3] such as non-constant rates and multi-lineage systems, we extend the mapping to a high-dimensional nonlinear function, instead of a conventional linear first-order equation in Equations 1 and 2.

Markov modeling of cell transitions

Existing works[1,2,38] usually leverage RNA velocities across genes to model cell transitions after solving gene-specific kinetic models, which treat RNA velocity estimation and cell transition modeling as two separate sequential parts. We argue that these two parts can be viewed as integrated ones, which is expected to boost the assessment of RNA velocity by virtue of cell transition modeling.

Under the common continuous assumption that cell states vary in small steps with numerous transitional populations,[1,38,39] we use a Markov Chain to model cell state transitions. In this approach, each state in the chain represents an observed cell profile, and edge weights indicate the likelihood of transitioning between two cells.

As RNA velocity is expected to accurately reflect cell transition dynamics at the local level, we leverage it to define transition probabilities of the Markov Chain. Unlike pseudo-time algorithms for trajectory inference, the randomness and the directed transition informed by RNA velocity should be considered, which are called diffusion and drift parts, as in previous work,[39] respectively.

The diffusion part aims to characterize the randomness of cell transitions, which restricts the possible transition direction in its neighborhood. In practice, the first step in chain construction is to compute an undirected kNN graph describing cell-cell similarities in a discrete manner. Formally, let $x_i \in R^M$ be the cellular state of cell $i \in \{1,2,\ldots,N\}$, and cell i's neighboring cells $j \in \{1,2,\ldots,N\}$, the diffusion kernel is defined as

$$d(x_i, x_j) = I\left(\frac{\|x_i - x_j\|}{r_k(x_i)}\right) \tag{Equation 4}$$

where I(·) is an indicator function with I(r) = 1 for $|r| \leq 1$ and 0 otherwise, and $r_k(x_i)$ is the distance from cell i to its k-th nearest neighbor. Besides, to better model cell-cell interaction, we further adopt Graph Attention (GAT) module on the kNN graph to capture the intercellular effect.

The drift part aims to depict the directed transition revealed by RNA velocity. Cell i is expected to have a high probability of transition toward cell j, when the direction of the truthful cellular state change $\delta_{ij} = x_j - x_i$ matches the predicted one by its RNA velocity v_i. Here, we choose the cosine scheme as the velocity kernel, i.e.,

$$r(x_i, x_j) = \cos\langle \delta_{ij}, v_i \rangle \tag{Equation 5}$$

where $\langle \cdot, \cdot \rangle$ is the angle between two input vectors. The overall transition kernel is then defined by,

$$k(x_i, x_j) = d(x_i, x_j) \cdot r(x_i, x_j) \tag{Equation 6}$$

The transition probability matrix $P = (p_{ij})_{i,j=1:N}$ among cells is finally defined by

$$p_{ij} = \frac{1}{z_i} \exp\left(\frac{k(x_i, x_j)}{\sigma_i^2}\right) \tag{Equation 7}$$

with row normalization factors $z_i = \sum_{j=1}^{N} \exp\left(\frac{k(x_i, x_j)}{\sigma_i^2}\right)$ and exponential kernel width parameters σ_i.

Temporal difference module

The objective of the temporal difference module is to estimate the continuous high-dimensional RNA velocity using Neural ODE in a bottom-up manner. The module takes an individual cell state vector $x \in R^{M \cdot d}$ as input and produces its corresponding first-order derivative $\dot{x} \in R^{M \cdot d}$, where RNA velocity is a part of the derivative.

Recent studies have shown that deep neural networks are effective in approximating high-dimensional nonlinear functions. Neural ODEs leverage the expressive power of neural networks to learn ODE from data without strict assumptions, making them suitable for our generalized RNA velocity model. However, the black-box nature of neural networks poses a limitation to their applications in the current problem, where biological interpretability is crucial in uncovering its mechanism.

Motivated by this, we design a tailored symbolic neural network (SymNet) to approximate the nonlinear mapping F in Equation 3, which has good expressive power and transparency. The analytic form of F can then be readily inferred after training (see below for more details). Specifically, we approximate the high-dimensional vector field F in Equation 3 by the SymNet F_θ, i.e.,

$$\dot{x} = F_\theta(x) \tag{Equation 8}$$

where θ is the parameters of neural networks. The pairs of temporally adjacent cells are used to supervise the optimization of this module. Given the present cell i, its corresponding future cell j is those whose state has the potential to transition from cell i in the Markov Chain, i.e.,

$$d(x_i, x_j) = 1 \text{ and } r(x_i, x_j) > \epsilon \tag{Equation 9}$$

where the first condition requires them to be connected in the kNN graph and the second one constrains the direction, $\epsilon > 0$ is a threshold to control the number of pairs. Here, we use pre-trained representation learning module[7] (an updated VeloAE model, see the "pre-trained representation learning module" section below for detail) to obtain the pairs, which implies the direction as a warm-up.

For the training of SymNet, cell pairs are selected based on the velocities calculated by the pre-trained VeloAE model. Considering a cell i with an expression state $x \in \mathbb{R}^{M \cdot d}$, and its neighboring cells $j \in \{1,2,\ldots,N\}$, the velocity v and spliced RNAs in latent space x^z, we select the pair (i, j) if the actual cellular state change direction from i to j aligns closely with the velocity v_i. Our selection method is as follows: with a 90% probability, we choose $j = \text{argmax}(\cos\langle v_i, x_j^z - x_i^z\rangle)$, and with a 10% probability, we randomly sample a neighboring cell as c_j, resulting in the pair (i, j).

Subsequently, we compile pairs for each cell and concatenate the cell state from each omics dataset to determine the true implied future cell state x^+. Throughout several epochs of training, we continually update these pairs. This iterative process ensures that both SymNet and VeloAE effectively learn from each other and accurately identify the correct cell velocities.

Furthermore, given the timescale φ, based on the forward Euler formula, the time difference loss for this branch is calculated as the difference between the predicted future cell state and the implied future cell state, as well as the regularization to ensure the sparsity of parameters, i.e.,

$$L_{TD} = \sum_{(x,x^+)} \|x + \varphi F_\theta(x) - x^+\|_2^2 + \lambda \|\theta\|_1 \tag{Equation 10}$$

where (x, x^+) is implied by a pre-trained representation learning module, and $x + \varphi F_\theta(x)$ is the predicted future cell state, $\lambda\|\theta\|_1$ is the sparse loss.

Symbolic neural network

Inspired by pde-net,[40] we develop a customized symbolic neural network called SymNet to approximate the continuous vector field. For simplicity, we first illustrate the SymNet in the gene-specific scenario, as shown in Figure S1A. The $SymNet_d^k$ is a network that takes a d dimensional vector as input and consists of k hidden layers. Each hidden layer of $SymNet_d^k$ directly takes the outputs from the preceding layer as inputs. Meanwhile, it introduces an extra variable (i.e., $f(\cdot,\cdot)$) at each hidden layer where f is a dyadic operation unit, e.g., multiplication or division. Here, we choose f as multiplication to increase the order, and the input to it are two linear combinations of the outputs from the preceding layer. The output of $SymNet_d^k$ is a linear combination of the outputs of the last hidden layer.

The $SymNet_d^k$ can represent all polynomials of input variables $x_1, x_2, \ldots, x_d$ with the total number of multiplications not exceeding k. Given the simplest scenario with two input variables u,s and no hidden layers, $SymNet_2^0(u, s)$ is actually a linear combination of two input variables, which is the conventional reaction equation in Equations 1 and 2.

Pre-trained representation learning module

The representation learning branch aims to learn a low-dimensional representation of RNA velocity that mitigates the effects of sparsity and noise present in raw counts, and it has demonstrated efficacy in revealing reliable cell transitions.[7] We adopt the main framework design in VeloAE, but replace the graph convolution (GCN) module[41] in the encoder with the graph attention (GAT) module.[42] The GAT module implicitly defines the weights of node aggregation by employing an attention mechanism over the node features, which is more congruous with cell interactions and performs better empirically than GCN module.

In contrast to the Temporal Difference branch, this branch accepts any transcriptome matrix $X \in R^{N \times M}$ as input, including spliced or unspliced reads (S or U), with normalized counts of M considered genes across N cells. After fitting the parameters of the autoencoder, we project S and U into the low-dimensional latent space using the encoder. All the velocities are then computed in a steady-state model.[2] We express the whole procedure of this branch as follows,

$$V_l = \mathcal{S}(\mathcal{E}(S), \mathcal{E}(U)) \in R^{N \times m} \tag{Equation 11}$$

where $\mathcal{E}(\cdot)$ is the encoder of this branch with GAT module as the main component, and $\mathcal{S}(\cdot,\cdot)$ denotes the operation of steady-state model,[2] which performs extreme-quantile regression on each column of latent matrix $\mathcal{E}(S)$ and $\mathcal{E}(U)$. Note that the dimension of velocity m is smaller than the corresponding velocity in the Temporal Difference branch M.

Mutual learning module

We adopt mutual learning[43] to further align the transition probabilities obtained from the two branches. Mutual learning,[43] a variant of knowledge distillation,[44] aims to start with a pool of untrained student networks that learn simultaneously to solve the task together. Through the collaborative optimization of each student network, they learn from each other throughout the learning process and inherit each other's strengths.

In our case, in addition to the intra-branch loss, each branch is trained with a mimicry loss that aligns with their respective transition probabilities. Let the transition probability matrix from TD and RL branch be $P_h, P_l \in R^{N \times N}$, respectively.

Further, for cell $i \in \{1,2,\ldots,N\}$, we denote its transition probability vectors from the two branches by $p_h^{(i)}, p_l^{(i)} \in R^N$. We then adopt the symmetric Jensen-Shannon Divergence loss as the mimicry loss. The objective to be minimized is defined as,

$$L_{mutual} = \sum_{i=1}^{N} \frac{1}{2}\left(D_{KL}\left(p_h^{(i)} \middle\| p_l^{(i)}\right) + D_{KL}\left(p_l^{(i)} \middle\| p_h^{(i)}\right)\right) \tag{Equation 12}$$

The procedure of mutual learning urges each branch to provide supervision for the other. The high-dimensional TD branch is expected to inherit reliable trajectories from the RL branch, while the RL branch acquires supervision for each cell in each latent dimension, unlike only extreme-quantile cells in VeloAE. During the inference phase, SymVelo primarily utilizes the SymNet branch to compute RNA velocity and cell differentiation trajectories, thereby leveraging the high-dimensional analysis capabilities of SymNet for enhanced accuracy and insight into cellular dynamics.

Process multimodal dataset

The 10x Multiome dataset is a multimodal dataset from human cerebral cortex. The cell state in the dataset includes unspliced mRNA abundance u, spliced mRNA abundance s, and chromatin accessibility c.

In this study, we address the challenge of extending the SymVelo model for efficient multimodal data analysis. Our solution involves enhancing the SymNet architecture and refining the loss function of the VeloAE model. $SymNet_d^k$, as described in Equation 3, processes an input $x \in \mathbb{R}^{M \cdot d}$, where d represents a flexible dimensionality corresponding to the number of modalities. This design offers adaptability for multimodal applications by minimizing the need for extensive modifications. Specifically, $SymNet_d^k$ can accurately represent all polynomials of the input variables x_1, x_2, …, x_d, with a constraint on the total number of multiplications not exceeding k. In the context of the 10x Multiome scenario, we set d as 3, considering three input variables (u, s, c) without incorporating hidden layers: $SymNet_3^0(u, s, c)$. Here, $SymNet_3^0(u, s, c)$ functions as a linear combination of these three variables, effectively capturing cell state as a concatenation of unspliced and spliced counts, along with chromatin accessibilities, $u, s, c \in R^M$, i.e., $x = (u, s, c)$.

SymNet can be flexibly extended to multimodal models in multimodal dataset, but the original VeloAE model cannot be extended to multimodal models. Therefore, during the network training process, there will be a mismatch with the dimension of VeloAE. To overcome this, we have enhanced VeloAE's encoder to efficiently reduce the dimensionality of multimodal cell state information, aligning it with the steady-state assumption prevalent in multi-omics problems. This enhancement is underpinned by the multimodal hypothesis outlined in MultiVelo[33] and is evident in Equations 13-15.

$$\frac{dc}{dt} = k_c\alpha_c - \alpha_c c(t) \tag{Equation 13}$$

$$\frac{du}{dt} = \alpha(t)c(t) - \beta u(t) \tag{Equation 14}$$

$$\frac{ds}{dt} = \beta u(t) - \gamma s(t) \tag{Equation 15}$$

where k_c indicates the open state of chromatin. When chromatin is accessible, $k_c = 1$ otherwise $k_c = 0$. α_c indicates chromatin rate parameter.

Moreover, we propose a refined loss function for VeloAE, as specified in Equations 16-18:

$$L_{rec} = MSE(s, \hat{s}) + MSE(u, \hat{u}) + MSE(c, \hat{c}) \tag{Equation 16}$$

$$L_{reg} = \sum_i MSE\left(c_i^z, \gamma_i \cdot u_i^z\right) \tag{Equation 17}$$

$$\mathrm{L} = L_{rec} + L_{reg} \tag{Equation 18}$$

where L_{rec} (reconstruction loss) calculated as mean squared error (MSE) between the reconstructed $\hat{s}$, $\hat{u}$, $\hat{c}$ and original s, u, c. The regression loss (L_{reg}) is computed on latent projection u^z and c^z to emulate the fitting of RNA degradation rate γ. In the low-dimensional space z, the degradation rate of the i th low dimension γ_i is fitted by solving the regression function $c = \gamma \cdot u + \theta$ on the extreme quantile cells at the i th dimension of u^z and c^z, respectively.

Ablation study of temporal difference module

To demonstrate the efficacy of SymVelo's Temporal Difference module, we conducted a detailed ablation study. The rationale behind this experimental design is to assess how effectively the Temporal Difference module processes cell pairs, specifically by analyzing the directional trajectory of cells based on their current state and their future states predicted by various models. We compared the cell pairs' current state as determined by the Temporal Difference module with their predicted future states obtained from three different models: VeloAE (as depicted in Figure 3A, right), scVelo (Figure S5C, left), and random modules (Figure S5C, right). Our study also included a quantitative assessment of Cell Behavior Directionality (CBDir) for each of these models. This metric evaluates the accuracy with which each model predicts the future state of cells. This analysis indicates that VeloAE is superior in accuracy for cell pairing.

Our methodology for selecting cell pairs involves a hybrid approach: a future cell predicted by VeloAE or scVelo is chosen 90% of the time, while a random neighboring cell from the cell proximity diagram is selected 10% of the time. Conversely, when utilizing the random module, our strategy diverges, involving the random selection of a neighboring cell from the cell proximity diagram to serve as the future cell in 100% of cases. This systematic and varied approach ensures a thorough validation of our model's predictive capabilities.

Pseudo-time inference

Root and endpoint are first determined by ''terminal_states'' function of scVelo. Then we use the ''velocity_pseudotime'' function of scVelo to infer pseudo-time based on velocity attained from SymVelo. All parameters are set to default.

Soft clustering of chromatin accessibility data

Data at the cellular level is extremely sparse, resulting in dramatic fluctuations even within the same cell type, while it is coarse at the cell type level. To mitigate this issue, we aggregate cells into pseudo-bulk samples by dividing cells into 10 groups along pseudo-time before soft clustering.[45] Mean values are calculated and standardized for each pseudo-bulk sample. To estimate the optimized number of cluster centroids c, we perform soft clustering with a range of cluster numbers from 2 to 20. And cluster number is determined by the centroid distance plot. We extract alpha cores of each cluster using acore = 0.5. Upon this procedure, we perform soft clustering on splicing, unsplicing and chromatin accessible features.

QUANTIFICATION AND STATISTICAL ANALYSIS

Cross-boundary direction correctness

Cross-boundary direction correctness (CBDir) serves as a quantitative metric for assessing the accuracy of the estimated velocity direction between different cell groups. It specifically evaluates the movement direction of cells across boundaries, necessitating input of established developmental directions between pairs of cell clusters, such as from Cell type A to Cell type B.

CBDir operates under the premise that both source and target cells are represented within a shared vector space. The developmental trajectory of a cell is approximated by computing the displacement from the source cell to the target cell. Consequently, for a predefined developmental direction from cell type A to type B, the ideal velocity of a type A cell should align with its displacement toward a type B cell. A positive estimated cell velocity indicates that the velocity direction is in harmony with the cell's developmental path. A higher CBDir value signifies greater alignment of the predicted direction with the true developmental direction. The equation of CBDir is:

$$\mathrm{CBDir}(c) = \frac{1}{|\{c' \in C_B \cap \mathcal{N}(c)\}|} \sum_{c' \in C_B \cap \mathcal{N}(c)} \frac{v_c \cdot (x_{c'} - x_c)}{|v_c| \cdot |x_{c'} - x_c|} \qquad \text{(Equation 19)}$$

where $x_{c'}$ and x_c are vectors representing cells c and c' in a low-dimensional space, as defined by the uniform manifold approximation and projection (UMAP). The term $x_{c'} - x_c$ represents cell displacement in this space, and v_c is d the decomposed UMAP velocity representation in the same space. Here, C_B denotes the set of type B cells, and $\mathcal{N}(c)$ symbolizes the neighboring cells of c.

Eta squared (correlation ratio)

Eta squared is a measure of effect size. We employ it here to evaluate the association between cell identity and velocity or RNA values, because in differentiated systems, cell transcription dynamics are often closely related to the evolution of cell types. The formula of Eta squared is the sum of squares of an effect (SS_{effect}) divided by the total sum of squares (SS_{total}). Higher values of Eta squared indicate a higher proportion of variance that a given variable can explain in the model.

$$\eta^2 = \frac{SS_{effect}}{SS_{total}} \qquad \text{(Equation 20)}$$

Skewness and spearman correlation

For each gene, we calculate their Velocity and RNA skewness using the "skewness" function in R. In addition, the Spearman correlation coefficient is calculated by "cor.test" function with setting method to "spearman".

Differential expression analysis and gene ontology enrichment

We use the "FindAllMakers" function of the Seurat[46] package to identify differential genes within each cell type (one versus others, p.adjust<0.01, log fold change>0.25). For the visualization of different sections in Figure 2F, we utilze the cutoff of log2FC > 0.5. Collection of gene sets (BP ontology) is used for over-presentation analysis by clusterProfiler.[47] We keep an ontology with smaller p value when the geneID is repeated and visualize top 10 ontology ordered by adjusted p values.