KNOWLEDGE-GUIDED RNA VELOCITY INFERENCE WITH LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding how cells transition between states is central to reconstructing dynamic biological processes from single-cell data. RNA velocity provides local, expression-derived estimates of transcriptional dynamics, but these signals do not explicitly encode the population-level biological structure governing plausible transitions between cell states. We therefore systematically investigate whether large language models (LLMs) can provide complementary biological reasoning for cell-state transition inference. Our benchmark reveals a clear capability boundary: LLMs are effective at identifying cell states and reasoning about state-level transitions, but are substantially less reliable at predicting fine-grained gene-level dynamics. Motivated by this complementarity, we introduce veloLLM, a knowledge-guided RNA velocity framework that preserves local expression-based dynamics while incorporating population-level transition priors to guide the organization of the velocity field. Across multiple single-cell datasets and base velocity estimators, veloLLM produces velocity fields that better align with known cell-state transitions while maintaining local dynamical coherence. Together, these results establish a principled division of labor between quantitative RNA velocity estimation and high-level biological reasoning, providing a general strategy for integrating biological knowledge into cell-state dynamics inference.

1 INTRODUCTION

Understanding how cells transition between states is fundamental to reconstructing dynamic biological processes from single-cell data. Single-cell RNA sequencing (scRNA-seq) provides high-resolution snapshots of cellular states but does not directly reveal how they evolve over time. RNA velocity (La Manno et al., 2018) addresses this limitation by leveraging unspliced and spliced transcripts to estimate the direction of cellular state changes. Subsequent methods, including dynamical models such as scVelo (Bergen et al., 2020) and probabilistic models such as veloVI (Gayoso et al., 2024), have further improved the estimation of transcriptional dynamics from static measurements.

Despite these advances, RNA velocity remains largely local and expression-driven. Existing methods infer short-range dynamics from gene expression, often using neighboring cells to construct transition probabilities or smooth velocity fields. Yet local transcriptional consistency does not guarantee globally plausible trajectories: different velocity models can yield substantially different directions on the same cellular manifold (Ancheta et al., 2026), and locally plausible transitions may conflict with known developmental or regulatory relationships. Cellular transitions are shaped not only by transcriptional similarity, but also by cell identity, developmental programs, regulatory interactions, and prior biological knowledge. Methods such as RegVelo (Wang et al., 2026) incorporate structured priors such as gene regulatory networks, but much relevant knowledge remains distributed across literature, annotations, databases, and heterogeneous biological descriptions.

This motivates a broader question: can large language models (LLMs) connect such biological knowledge with quantitative cell-state dynamics? Recent work has begun adapting LLMs to single-cell biology. Cell2Sentence represents ranked expression profiles as text-like "cell sentences" (Levine et al., 2024), while CellTok models transcriptomes as discrete tokens jointly with textual instructions and biological context (Xiao et al., 2025). VCWorld further combines structured biological knowledge with iterative LLM reasoning to simulate perturbation responses and generate mechanistic explanations (Wei et al., 2026). Benchmarks suggest that LLMs can reliably annotate

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

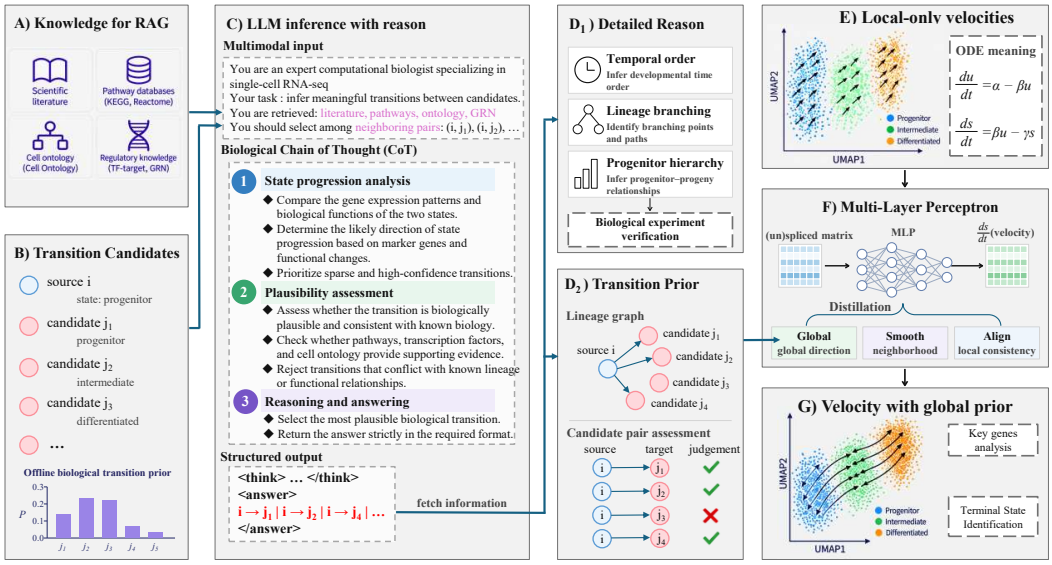


Figure 1: Overview of veloLLM, which uses retrieval-augmented LLM reasoning to construct global transition priors for RNA velocity refinement. (A) Scientific literature, pathway databases, cell ontologies, and regulatory knowledge form the retrieval corpus. (B) Neighboring cell-state pairs define candidate transitions. (C) Conditioned on the retrieved evidence, the LLM reasons over state progression and biological plausibility to select directed transitions. Its structured predictions provide both (D_1) interpretable rationales about temporal order, lineage branching, and progenitor hierarchy and (D_2) a sparse, directed transition prior. (E) A base kinetic model independently estimates local cell-wise velocities from splicing dynamics. (F) A multilayer perceptron refines these velocities by distilling the LLM prior through a global-direction objective, together with neighborhood smoothness and local-consistency constraints. (G) The resulting velocity field preserves local kinetic evidence while gaining globally coherent, biologically grounded directions for downstream analysis.

broad cell types from marker genes (Crowley et al., 2025), but remain less reliable for fine-grained response and perturbation tasks (Zhang et al., 2025), with unsupported biological claims still posing a risk (Xie et al., 2026). Cell-state transition inference therefore remains challenging: it requires both coarse-grained knowledge of cell identities and developmental relationships and fine-grained quantitative modeling of expression changes. Can these complementary strengths be combined to model cellular transitions more faithfully?

To answer this question, we systematically benchmark LLMs for cell-state transition reasoning across tasks that range from cell-state identification and semantic alignment to transition prediction, gene-level classification, expression-direction prediction, and gene prioritization. The results reveal a clear capability boundary: LLMs identify cell states, align heterogeneous biological semantics, recognize plausible developmental relationships, and reason about qualitative gene-regulation trends, but struggle with fine-grained ranking and quantitative estimation in high-dimensional gene-expression space. These findings suggest that LLMs are better suited to providing biological context and global priors than to directly replacing quantitative RNA velocity inference.

This capability boundary motivates a complementary design in which expression-based velocity models preserve fine-grained, quantitative local dynamics, while LLMs align cell-state descriptions that vary across datasets in nomenclature and granularity (Hu et al., 2025) and provide global biological context. We therefore introduce a knowledge-guided RNA velocity framework (Figure 1) that organizes heterogeneous prior knowledge into semantically aligned transition priors and integrates them with quantitative RNA velocity to guide global directions. By using biological knowledge to determine which locally supported transitions are globally plausible, the framework retains cell-level transcriptomic evidence while yielding velocity fields that are more biologically coherent and interpretable in terms of relevant cell identities, genes, and regulatory processes.

Our contributions are twofold:

- **Systematic characterization of LLM capabilities for cell-state transition reasoning.** We characterize the strengths and limitations of LLMs in cell-state identification, semantic alignment, and qualitative biological reasoning, showing that they remain unreliable for direct fine-grained transcriptional dynamics estimation. These findings motivate a division of labor in which LLMs provide high-level biological reasoning and quantitative velocity estimators capture local transcriptional dynamics.
- **Knowledge-guided RNA velocity with global biological constraints.** We develop veloLLM, a framework that translates heterogeneous biological knowledge into computational priors through cell-state semantic alignment and plausible transition constraints. Across multiple single-cell datasets and base velocity estimators, integrating these priors with local expression-based estimates improves alignment with known cell-state transitions while maintaining local dynamical coherence. These results demonstrate how global biological knowledge can complement local quantitative estimation to guide cell-state dynamics inference.

2 Related Work

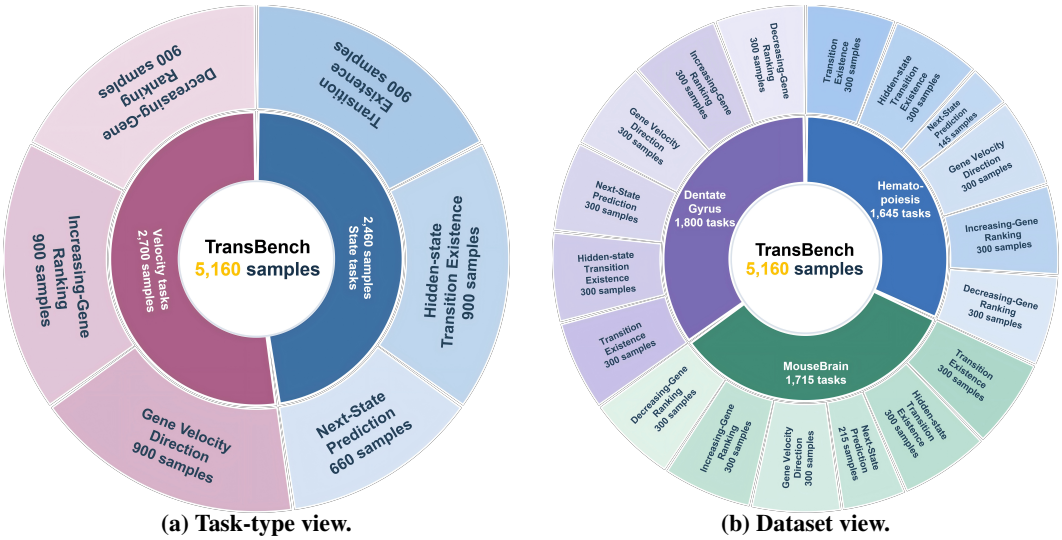
Cell State Transition Modeling. Modeling cell state transitions through development, differentiation, and perturbation responses is a central challenge in computational biology. RNA velocity infers transition directions from post-transcriptional splicing kinetics, with scVelo (Bergen et al., 2020) establishing the biophysical foundation under steady-state or full-dynamical assumptions. veloVI (Gayoso et al., 2024) and DeepVelo (Cui et al., 2024) relax shared kinetic-rate assumptions to capture cell-specific heterogeneity, yet rely solely on observed expression with regularization limited to local neighborhood smoothing. Geometric Flow Grounding (GFG) encodes data manifold geometry through Neural Tangent Projection, confining velocities to the tangent bundle to eliminate off-manifold hallucinations; however, this constraint derives solely from observed distributions without external biological semantics (Yu et al., 2026). RegVelo (Wang et al., 2026) incorporates external knowledge through pre-defined gene regulatory networks (GRNs), but its reliance on dataset-specific, fixed GRNs limits flexibility and cross-dataset generalization across varying cell-state annotations and biological contexts. These limitations motivate a more flexible way to translate heterogeneous biological knowledge into global constraints on cell-state transitions.

Large Language Models for Single-Cell Biology. Recent studies have increasingly explored large language models (LLMs) as a unified framework for representing and reasoning about single-cell biology. Cell2Sentence (C2S) and its scaled-up C2S-Scale models convert transcriptomic profiles into ranked-gene sequences for cell-state interpretation and perturbation-response prediction (Levine et al., 2024). CellTok (Xiao et al., 2025) explores token-based generative representations of cells, while CellWhisperer (Schaefer et al., 2025) aligns transcriptomic states with natural-language descriptions for semantic retrieval and annotation. OmicsLM (Sypetkowski et al., 2026) extends LLM-based modeling to broader omics understanding and zero-shot perturbation reasoning, and PerturbQA (Wu et al., 2025a) benchmarks perturbation reasoning and introduces a retrieval-augmented framework for predicting cellular responses. These studies connect cellular data with biological knowledge, but using that knowledge to guide continuous cell-state dynamics remains underexplored. veloLLM addresses this gap by organizing knowledge from ontologies and literature into semantically aligned transition priors that softly guide quantitative RNA velocity.

3 Preliminaries

Problem Formulation Given a single-cell RNA sequencing dataset with N cells and G genes, let $\mathbf{U}, \mathbf{S} \in \mathbb{R}^{N \times G}$ denote the observed unspliced and spliced expression matrices. Their entries u_{ig} and s_{ig} give the abundances of gene g in cell i, and $\mathbf{u}_i, \mathbf{s}_i \in \mathbb{R}^G$ are the corresponding cell-wise vectors. Under the classical RNA velocity model, the dynamics of gene g follow

$$\frac{du_g(t)}{dt} = \alpha_g - \beta_g u_g(t), \qquad v_g(t) = \frac{ds_g(t)}{dt} = \beta_g u_g(t) - \gamma_g s_g(t),$$



(a) Task-type view. **(b) Dataset view.**

Figure 2: Scale and task coverage of TransBench. The task-type view (a) summarizes the six evaluation tracks, while the dataset view (b) shows the corresponding coverage for Hematopoiesis, MouseBrain, and DentateGyrus.

where α_g, β_g, and γ_g denote the gene-specific transcription, splicing, and degradation rates, respectively.

Based on these kinetics, a base estimator provides an initial velocity field $\mathbf{V} = (v_{ig})_{i,g} \in \mathbb{R}^{N\times G}$, with cell-wise velocity vectors $\mathbf{v}_i \in \mathbb{R}^G$. To guide the direction of this field, we introduce a population-level transition prior $\mathcal{E}_{\text{prior}} \subseteq \mathcal{C} \times \mathcal{C}$, where $\mathcal{C}$ is the set of annotated cell states and each directed edge specifies a proposed source-to-target transition. Given the expression profile $\mathbf{x}_i \in \mathbb{R}^G$ of cell i, veloLLM learns a refiner

$$\hat{\mathbf{v}}_i = f_\theta(\mathbf{x}_i), \qquad \hat{\mathbf{V}} = (\hat{v}_{ig})_{i,g} \in \mathbb{R}^{N\times G}.$$

The base velocity field and transition prior supervise the refiner through velocity-alignment and edge-guidance losses, together with a within-state smoothness constraint. The resulting field $\hat{\mathbf{V}}$ integrates local kinetic estimates with population-level biological guidance.

4 Method

4.1 TransBench: A Cell State Transition Benchmark

TransBench scope. We introduce TransBench, a benchmark for evaluating whether large language models (LLMs) can reason about local cell-state transitions and their associated gene changes.

Datasets and task coverage. TransBench covers three developmental systems: human hematopoietic differentiation (Hematopoiesis), adolescent mouse nervous system (MouseBrain), and postnatal mouse dentate-gyrus neurogenesis (DentateGyrus). Together they contain 8,242 cells and 5,160 task instances. Figure 2 shows the same coverage in two complementary views: panel (a) groups instances by task type, whereas panel (b) groups them by dataset.

Tasks. We evaluate six complementary tracks: Transition Existence (TE), Hidden-state Transition Existence (HTE), Next-State Prediction (NSP), Gene Velocity Direction (GVD), Increasing-Gene Ranking (IGR), and Decreasing-Gene Ranking (DGR). The first three assess cell-state transitions, whereas the latter three assess gene-level directional changes and rankings. Dataset harmonization and sample construction are described in the appendix.

TransBench evaluation. We compare four controlled settings: P1 uses neither retrieval nor reasoning, P2 adds biological RAG, P3 enables reasoning, and P4 combines both. We evaluate

Table 1: Performance of four LLMs and the matched RNA-velocity baselines on TransBench. Each dataset is shown as a separate group row, followed by the model results. TE, HTE, NSP, and GVD report Macro-F1; IGR and DGR report NDCG@20. Higher is better for every column.

Model / Method	Classification				Ranking	
	TE	HTE	NSP	GVD	IGR	DGR
Hematopoiesis						
Qwen3.8-Flash	.922/.989/.816/.913	.564/.583/.515/.592	.911/**1.000**/.643/.821	.529/.507/.501/.522	.144/.150/.164/.160	.162/.163/.172/.172
DeepSeek-V4.1-Flash	.974/**.996**/.835/.908	.541/.577/.551/**.724**	.417/.674/.873/.902	.483/.492/.510/.543	.166/.148/.160/.157	.168/.187/.200/.195
Kimi-K2.6	.881/**.996**/.872/.977	.571/**.611**/.527/.579	.895/.999/.775/.999	.484/.529/.526/**.568**	.139/.129/.137/.130	.150/.184/.149/.171
GPT-5.6-Sol	.956/.800/.806/.985	.533/.512/.534/.536	.542/.493/.528/.490	.482/.536/.566/.568	.152/.156/.158/.159	.181/.228†/.218/.200
RegVelo-derived baseline	–	–	.286	.557	**.209**	**.294**
MouseBrain						
Qwen3.8-Flash	.808/.974/.795/.863	.749/.762/.669/.682	.577/.374/.377/.496	.606/.609/.604/.531	.192/.189/.192/.182	.261/.261/.256/.240
DeepSeek-V4.1-Flash	.744/.967/.993/**1.000**	.553/.573/.752/.757	.283/.297/.575/.534	.502/.511/.609/.580	.149/.188/.210/**.218**	.281/.245/.230/.239
Kimi-K2.6	.638/.750/.663/.705	.556/.659/.573/.653	.454/**.693**/.330/.626	.572/.584/.550/.518	.168/.176/.164/.179	.257/.270/.238/.241
GPT-5.6-Sol	.858/.971/.963/.874	**.842**/.811/.754/.769	.223/.427/.404/.371	**.644**/.635/.601/.626	.204/.218/.218/.207	.248/.269/.262/.258
DeepVelo-derived baseline	–	–	.090	.630	.161	**.282**
DentateGyrus						
Qwen3.8-Flash	.985/**1.000**/.921/.953	.614/.576/.571/.622	.887/**1.000**/.888/.927	.497/.522/.502/.509	.183/.190/.186/.179	.170/.182/.166/.171
DeepSeek-V4.1-Flash	.965/.989/**1.000/1.000**	.561/.559/.502/.581	.769/.930/**1.000/1.000**	.477/.490/.574/.550	.195/.203/.220/.218	.172/.190/.204/.197
Kimi-K2.6	.954/.982/.881/.937	.542/.559/.535/.571	.979/.993/.857/**1.000**	.573/.610/.533/.545	.200/.196/.196/.203	.161/.168/.154/.166
GPT-5.6-Sol	**1.000/1.000/1.000/1.000**	.667/.635/.632/**.691**	.985/.996/.970/**1.000**	.621/.603†/.603/**.629**	**.236**/.230/.229/.228	.213/**.219**/.207/.208
veloVI-derived baseline	–	–	.252	.516	.199	.186

Notes: Each LLM cell lists P1/P2/P3/P4 in order, where P1=(RAG off, reasoning off), P2=(RAG on, reasoning off), P3=(RAG off, reasoning on), and P4=(RAG on, reasoning on). A dash indicates that a velocity baseline is not applicable to TE or HTE. Pale-purple rows denote the matched RNA-velocity baselines. † marks a partially evaluated result: Hematopoiesis GPT-P2 DGR ($n = 297$) and DentateGyrus GPT-P2 GVD ($n = 298$). All other LLM entries use the complete dataset-specific TransBench task set: $n = 300$ per task except Hematopoiesis NSP ($n = 145$) and MouseBrain NSP ($n = 215$). Bold and underlined values are the best and second-best complete results, respectively, within each dataset and task; partial results are excluded from this highlighting.

Qwen3.8-Flash, DeepSeek-V4.1-Flash, Kimi-K2.6, and GPT-5.6-Sol using Macro-F1 for TE, HTE, NSP, and GVD, and NDCG@20 for IGR and DGR. The appendix provides the full prompt templates, controls, and evaluation details.

Results and analysis. Table 1 reports all currently available full-scale results. Classification columns report Macro-F1 and ranking columns report NDCG@20.The results show a consistent granularity gap: LLMs perform strongly on coarse cell-state transition tasks, but accuracy drops on velocity-oriented gene-level direction and ranking tasks, where they are often below the matched baselines. LLMs therefore provide useful high-level biological guidance for RNA-velocity analysis, but do not replace dedicated velocity models for fine-grained dynamic prediction.

4.2 Incorporating LLM Priors into Base Model Predictions

veloLLM refines RNA velocity estimation by coupling local kinetic dynamics with population-level priors inferred by an LLM (Figure 1). Given cell expression profiles and population annotations, task-aware retrieval assembles evidence from curated gene annotations, cell ontologies, observed gene programs, and independent literature, following Section 4.1. The LLM uses this evidence to align heterogeneous state names and predict directed source–target transitions, which define the population-level edge prior $\mathcal{E}_{\text{prior}}$ rather than cell-level velocities. In parallel, an existing RNA velocity estimator provides a reference velocity field used for kinetic alignment: instead of directly serving as the final prediction, its velocity estimates distill the local dynamical semantics implied by the underlying RNA-velocity ODEs into the refiner. Specifically, a refiner f_θ maps each expression profile $\mathbf{x}_i$ to $\hat{\mathbf{v}}_i = f_\theta(\mathbf{x}_i)$, and is trained to preserve these ODE-consistent local dynamics while incorporating population-level transition constraints. Training therefore combines velocity-alignment distillation, within-state smoothness, and the LLM-guided edge loss; the architecture and loss definitions are given in Appendix A.4.

LLM-guided edge-prior loss. Each directed edge $(u, v) \in \mathcal{E}_{\text{prior}}$ identifies a proposed transition from source state u to target state v. We apply guidance only to source cells with at least

Table 2: Computation of the LLM-guided edge-prior loss. The directed prior selects source–target populations; all directions are computed in gene-expression space.

Input: Expression vectors $\mathbf{x}_i$, predicted velocities $\hat{\mathbf{v}}_i$, state labels c_i, neighbor sets $\mathcal{N}(i)$, and edge prior $\mathcal{E}_{\text{prior}}$; $\tau = 0.5$, $T = 0.1$.

1. Set $\mathcal{Q} = \emptyset$. For each $(u, v) \in \mathcal{E}_{\text{prior}}$, visit source cells i with $c_i = u$.
2. Set $\mathcal{N}_i^v = \{j \in \mathcal{N}(i) : c_j = v\}$; skip i if $\mathcal{N}_i^v = \emptyset$.
3. For each $j \in \mathcal{N}_i^v$, compute $\mathbf{\Delta}_{ij} = \mathbf{x}_j - \mathbf{x}_i$ and $s_{ij} = \mathrm{n}(\hat{\mathbf{v}}_i)^\top \mathrm{n}(\mathbf{\Delta}_{ij})$.
4. Weight target neighbors by $\alpha_{ij} = \exp(s_{ij}/\tau) / \sum_{k \in \mathcal{N}_i^v} \exp(s_{ik}/\tau)$.
5. Form $\bar{\mathbf{d}}_i = \mathrm{n}\Big(\sum_{j \in \mathcal{N}_i^v} \alpha_{ij} \mathbf{\Delta}_{ij}\Big)$ and append $\ell_{uvi} = \big(1 - \mathrm{n}(\hat{\mathbf{v}}_i)^\top \bar{\mathbf{d}}_i\big)^2$ to $\mathcal{Q}$.
6. If $\mathcal{Q} = \emptyset$, return 0; otherwise set $\omega_q = \exp(\mathrm{sg}(\ell_q)/T) / \sum_{r \in \mathcal{Q}} \exp(\mathrm{sg}(\ell_r)/T)$.
Output: $\mathcal{L}_{\text{edge}} = \sum_{q \in \mathcal{Q}} \omega_q \ell_q$, where $\mathrm{n}(\mathbf{a}) = \mathbf{a} / \max(\|\mathbf{a}\|_2, \varepsilon)$ and sg stops gradients.

one target-state neighbor, using their gene-expression displacements as candidate transition directions. A velocity-dependent softmax combines these directions, and the resulting directional error is weighted across source cells so that poorly aligned cells contribute more to training. Table 2 details both weighting stages.

5 Experiments

5.1 Experimental Setup

Datasets and Baselines. We evaluate veloLLM on four single-cell datasets spanning different developmental systems: *MouseBrain* (Zeisel et al., 2018), *Hematopoiesis* (Qiu et al., 2022), *Dentate Gyrus* (Hochgerner et al., 2018), and *Neural Crest* (Kastriti et al., 2022). We use five RNA velocity estimators as baselines: *scVelo* (Bergen et al., 2020), *veloVI* (Gayoso et al., 2024), *DeepVelo* (Cui et al., 2024), *RegVelo* (Wang et al., 2026), and *Geometric Flow Grounding* (Yu et al., 2026). For each estimator, we compare its original velocity field with the field refined by veloLLM on the same datasets, using matched evaluation metrics and protocols wherever applicable.

Evaluation Metrics. We assess velocity-field quality using Velocity Coherence (VeloCoh), Intra-Cluster Coherence (ICCoh), and Cross-Boundary Direction (CBdir). VeloCoh and ICCoh measure local and within-cluster velocity coherence, respectively, while CBdir measures directional alignment with reference cell-state transitions.

Implementation Details. We train the velocity refiner using Adam with a learning rate of 1×10^{-4} and weight decay of 1×10^{-5}. The three loss terms use fixed weights of $\lambda_{\text{mse}} = 1$, $\lambda_{\text{smooth}} = 1$, and $\lambda_{\text{edge}} = 1$ throughout training. The model is trained for 3000 epochs by default.

5.2 Comparison Results

In the main comparison in Table 3, veloLLM refines the base velocity fields using dataset-specific transition priors informed by the literature. CBdir improves across all base-model–dataset comparisons, while ICCoh increases in most cases and changes in VeloCoh depend on the baseline. For example, veloVI gains in both coherence metrics on Neural Crest, whereas GFG, which has the highest baseline coherence across datasets, gains in CBdir despite reduced coherence on Dentate Gyrus. These trends suggest that veloLLM improves developmental direction while balancing directional guidance with local organization. The edge prior favors source-to-target displacements when source cells have target-state neighbors, while the base-velocity reconstruction and within-state smoothness losses constrain changes to the field. This balance allows refinement to strengthen local coherence or relax an already smooth field when improving its biological alignment. We separately evaluate the effect of transitional priors generated by language models on veloLLM in Appendix A.6.

Table 3: Quantitative comparison across four scRNA-seq datasets. Velocity Coherence (VC) measures local agreement among neighboring-cell velocities, Intra-Cluster Coherence (ICC) measures velocity-direction agreement within annotated clusters, and Cross-Boundary Direction (CBD) measures alignment with annotated source-to-target transitions at cluster boundaries; higher is better for all metrics. $\Delta = (\text{Our} - \text{Base}) \times 100\%$ denotes the gain of veloLLM. Bold denotes the highest mean for each dataset and metric.

Method	Mouse Brain			Hematopoiesis			Dentate Gyrus			Neural Crest		
	VC	ICC	CBD	VC	ICC	CBD	VC	ICC	CBD	VC	ICC	CBD
scVelo	0.779 ±0.000	0.530 ±0.000	0.494 ±0.062	0.682 ±0.000	0.511 ±0.000	-0.311 ±0.026	0.816 ±0.010	0.674 ±0.013	0.532 ±0.151	0.805 ±0.043	0.609 ±0.009	-0.091 ±0.043
Our (scVelo)	0.865 ±0.001	0.639 ±0.001	0.792 ±0.041	0.675 ±0.000	0.814 ±0.000	0.366 ±0.037	0.767 ±0.041	0.814 ±0.036	0.747 ±0.016	0.919 ±0.003	0.775 ±0.007	0.696 ±0.005
Δ (%)	+8.6%	+10.9%	+29.8%	-0.7%	+30.3%	+67.7%	-4.9%	+14.0%	+21.5%	+11.4%	+16.6%	+78.7%
veloVI	0.826 ±0.001	0.601 ±0.002	0.368 ±0.015	0.749 ±0.017	0.645 ±0.018	0.121 ±0.021	0.835 ±0.006	0.651 ±0.014	0.702 ±0.022	0.773 ±0.004	0.508 ±0.008	0.087 ±0.044
Our (veloVI)	0.946 ±0.000	**0.938** ±0.000	0.788 ±0.038	0.748 ±0.000	0.864 ±0.000	0.385 ±0.011	0.908 ±0.000	0.829 ±0.000	0.854 ±0.019	0.924 ±0.000	0.877 ±0.000	0.591 ±0.046
Δ (%)	+12.0%	+33.7%	+42.0%	-0.1%	+21.9%	+26.4%	+7.3%	+17.8%	+15.2%	+15.1%	+36.9%	+50.4%
DeepVelo	0.766 ±0.004	0.349 ±0.003	0.380 ±0.193	0.888 ±0.023	0.717 ±0.013	0.329 ±0.036	0.917 ±0.004	0.817 ±0.006	0.265 ±0.098	0.862 ±0.002	0.469 ±0.014	0.362 ±0.038
Our (DeepVelo)	0.875 ±0.005	0.662 ±0.020	**0.823** ±0.004	0.675 ±0.003	0.798 ±0.006	0.539 ±0.010	0.841 ±0.001	0.871 ±0.000	0.848 ±0.001	0.916 ±0.003	0.867 ±0.003	**0.797** ±0.004
Δ (%)	+10.9%	+31.3%	+44.3%	-21.3%	+8.1%	+21.0%	-7.6%	+5.4%	+58.3%	+5.4%	+39.8%	+43.5%
RegVelo	0.953 ±0.004	0.931 ±0.003	-0.791 ±0.004	0.855 ±0.001	0.703 ±0.005	0.346 ±0.043	0.894 ±0.007	0.771 ±0.013	0.103 ±0.014	0.921 ±0.000	0.734 ±0.000	-0.126 ±0.000
Our (RegVelo)	0.928 ±0.000	0.914 ±0.000	0.796 ±0.001	0.716 ±0.000	0.809 ±0.000	**0.632** ±0.005	0.757 ±0.000	0.764 ±0.000	0.686 ±0.000	0.885 ±0.000	0.862 ±0.000	0.681 ±0.000
Δ (%)	-2.5%	-1.7%	+158.7%	-13.9%	+10.6%	+28.6%	-13.7%	-0.7%	+58.3%	-3.6%	+12.8%	+80.7%
GFG	**0.980** ±0.001	0.936 ±0.002	0.580 ±0.061	**0.976** ±0.017	**0.962** ±0.028	0.033 ±0.034	**0.967** ±0.004	**0.937** ±0.006	0.406 ±0.143	**0.981** ±0.005	**0.940** ±0.021	0.436 ±0.044
Our (GFG)	0.960 ±0.002	0.915 ±0.002	0.727 ±0.087	0.684 ±0.013	0.818 ±0.007	0.398 ±0.079	0.880 ±0.013	0.854 ±0.003	**0.893** ±0.007	0.916 ±0.001	0.855 ±0.007	0.716 ±0.019
Δ (%)	-2.0%	-2.1%	+14.7%	-29.2%	-14.4%	+36.5%	-8.7%	-8.3%	+48.7%	-6.5%	-8.5%	+28.0%

5.3 Ablation Study

Component Ablation. We ablate the three loss components in veloLLM + DeepVelo (Figures 3 and 4). Removing the edge prior weakens transition directionality despite slightly improved velocity coherence, suggesting that a locally smooth field can still follow an incorrect developmental trajectory. Removing smoothness reduces local and within-cluster coherence while largely preserving directionality, yielding a less coordinated velocity field. Without MSE alignment (*w/o ODE*), local ODE scores decline across much of the map, and velocities of more neighboring cells deviate from the u–s plane, weakening their consistency with RNA kinetics. The three terms thus provide complementary constraints on transition direction, local coherence, and kinetic consistency; directional plausibility alone does not ensure a biologically meaningful velocity field.

Parameter Sensitivity. We examine the sensitivity to the three loss weights on the *Hematopoiesis* dataset using scVelo as the base model (Table 4). A smaller reconstruction weight yields the best overall performance, while overly strong reconstruction regularization degrades the results, suggesting that excessive adherence to the base velocity can limit refinement. Increasing the smoothness weight mainly improves local coherence with little effect on transition direction, confirming that this term primarily controls neighborhood consistency. In contrast, increasing the edge weight substantially improves directional alignment while only moderately affecting coherence, indicating that the edge prior mainly governs biologically aligned transition direction.

5.4 Terminal States Identification

We evaluate terminal-state identification on *Hematopoiesis* using GPCCA and the TSI score from CellRank 2 (Weiler et al., 2024), with implementation details and additional comparisons in Appendix A.7. As shown in Figure 5, veloLLM improves terminal-state recovery at low-to-

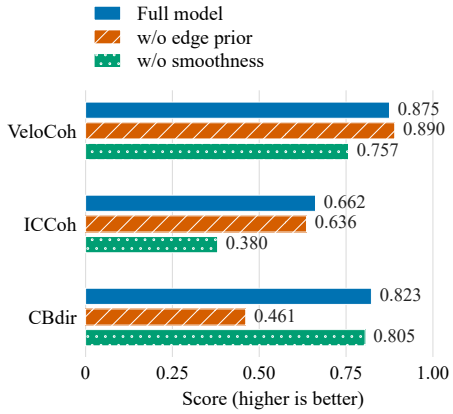
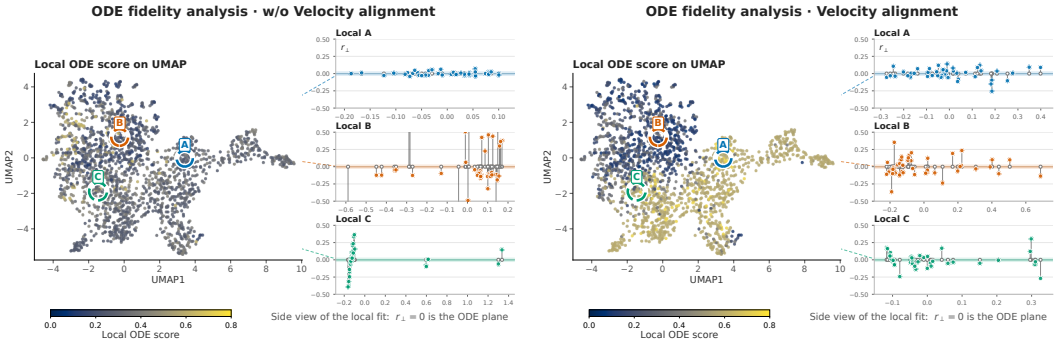


Figure 3: Component ablation.

Table 4: Parameter sensitivity analysis

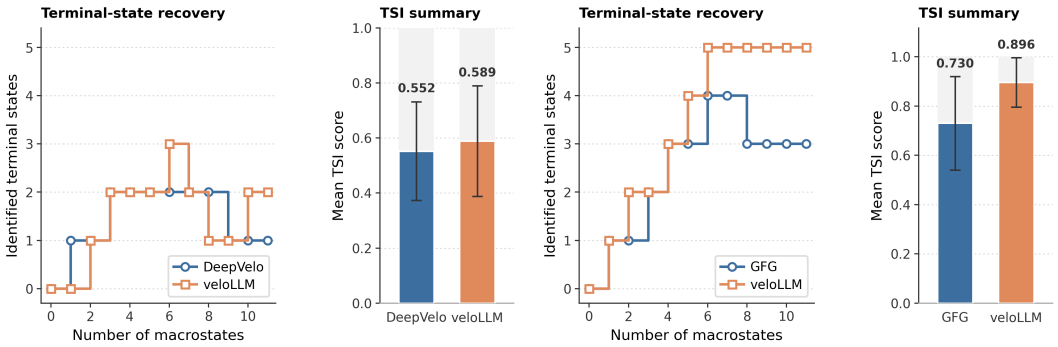
Parameter	Value	VeloCoh	ICCoh	CBdir
λ_{mse}	0.1	0.912 ±0.000	0.886 ±0.000	0.827 ±0.001
	1	0.865 ±0.000	0.640 ±0.001	0.773 ±0.001
	10	0.831 ±0.001	0.585 ±0.002	0.501 ±0.001
λ_{smooth}	0.1	0.825 ±0.000	0.583 ±0.002	0.768 ±0.002
	1	0.865 ±0.000	0.640 ±0.001	0.773 ±0.001
	10	0.944 ±0.000	0.884 ±0.001	0.768 ±0.001
λ_{edge}	0.1	0.871 ±0.000	0.641 ±0.000	0.508 ±0.006
	1	0.865 ±0.000	0.640 ±0.001	0.773 ±0.001
	10	0.834 ±0.001	0.651 ±0.000	0.835 ±0.000



(a) w/o velocity alignment

(b) velocity alignment

Figure 4: Local ODE fidelity of the refined velocity. Left panels show per-gene, neighborhood-level ODE fit scores on the same UMAP (higher is better), with a shared representative gene across both runs. Right subplots show side views of each region's fitted plane in centered, robustly scaled unspliced–spliced–velocity space. The horizontal zero line marks the local ODE plane; colored points have signed orthogonal residuals $r_\perp$, and vertical segments indicate their distances to the plane.



(a)

(b)

Figure 5: Terminal-state recovery and mean TSI on Hematopoiesis using (a) DeepVelo and (b) GFG as base velocity models. Terminal-state recovery is evaluated across increasing macrostate resolutions for the original and veloLLM-refined velocity fields.

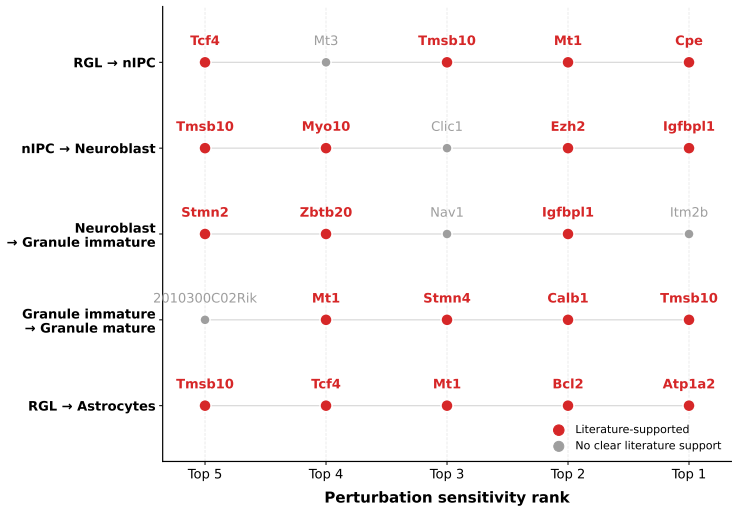


Figure 6: Top five perturbation-sensitive genes for selected Dentate Gyrus transitions. Red indicates literature-supported associations with the relevant cell identity, lineage, or differentiation process; gray indicates genes without clear literature support.

intermediate macrostate resolutions compared with DeepVelo, while combining veloLLM with GFG enables earlier recovery of all annotated endpoints and preserves them at higher resolutions. Consistently, TSI improves with both DeepVelo and GFG, indicating that knowledge-guided refinement enhances overall terminal-state identification and robustness across resolutions.

5.5 Perturbation-Sensitive Genes in Cell-State Transitions

Figure 6 shows the top five genes ranked by perturbation-induced changes in velocity direction for selected Dentate Gyrus transitions. Most identified genes have established biological associations with the corresponding cell states or developmental processes, spanning progenitor migration and proliferation, early neuronal maturation, and mature granule-cell identity. Appendix A.8 provides the scoring procedure and detailed biological analysis.

6 Conclusion

Our benchmark identifies complementary roles for LLMs and expression-based models in cell-state transition inference: LLMs reason effectively about relationships between annotated cellular states, whereas fine-grained transcriptional dynamics remain better grounded in quantitative expression signals. Building on this insight, we developed veloLLM, a plug-in framework that refines a base RNA velocity field through population-level transition priors, velocity alignment, and within-state smoothness. With literature-informed priors, veloLLM improved CBdir in all 20 comparisons across four datasets and five base models, and ICCoh in 14. Ablations further show that transition guidance, local coherence, and alignment with the base velocity serve distinct roles in refinement. Together, these findings support combining biological knowledge with local kinetic estimates to improve the direction of inferred cell-state transitions.

Limitations

veloLLM depends on transition-prior quality, making improved prior accuracy and coverage an important direction for future work. Our experiments integrate and distill heterogeneous biological knowledge for RNA velocity inference, but the incorporated priors still capture only part of the knowledge relevant to cell-state transitions. Future work could incorporate broader priors and explore their potential for modeling more diverse cell-state dynamics.

AI USE STATEMENT

In this work, we used generative AI tools to assist with literature search and to improve the clarity, grammar, and presentation of the manuscript. In the context of research ideation or execution, the use of generative AI was limited to supporting literature discovery and related exploratory research activities. All references identified with AI assistance were independently verified by the authors against the original sources.

Generative AI tools were not used to develop the conceptual framework, formulate hypotheses, design the methodology or experiments, conduct qualitative analysis, interpret experimental results, or generate synthetic data. These tasks were performed independently by the authors or were not applicable to this work.

All AI-assisted outputs and edits were critically reviewed and, where necessary, revised by the authors. The authors take full responsibility for the final content of this work, including all text, claims, analyses, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The problem formulation and method are described in the main paper, including the model components and training objectives. The experimental section reports the datasets, evaluation tasks, metrics, baselines, and ablations used in our evaluation, and the appendix provides the model-scaling analysis.

REFERENCES

Sarah Ancheta, Leah Dorman, Guillaume Le Treut, Abel Gurung, Greg Huber, Loïc A Royer, Alejandro Granados, and Merlin Lange. Challenges and progress in rna velocity: Comparative analysis across multiple biological contexts. *PLOS Computational Biology*, 22(6):e1014303, 2026.

Volker Bergen, Marius Lange, Stefan Peidli, F Alexander Wolf, and Fabian J Theis. Generalizing rna velocity to transient cell states through dynamical modeling. *Nature biotechnology*, 38(12): 1408–1414, 2020.

George Crowley, Tabula Sapiens Consortium, and Stephen R. Quake. Benchmarking cell type and gene set annotation by large language models with AnnDictionary. *Nature Communications*, 16: 9511, 2025. doi: 10.1038/s41467-025-64511-x.

Haotian Cui, Hassaan Maan, Maria C Vladoiu, Jiao Zhang, Michael D Taylor, and Bo Wang. Deepvelo: deep learning extends rna velocity to multi-lineage systems with cell-specific kinetics. *Genome Biology*, 25(1):27, 2024.

Adam Gayoso, Philipp Weiler, Mohammad Lotfollahi, Dominik Klein, Justin Hong, Aaron Streets, Fabian J Theis, and Nir Yosef. Deep generative modeling of transcriptional dynamics for rna velocity analysis in single cells. *Nature methods*, 21(1):50–59, 2024.

Chenying Guo, Kin-Sang Cho, Yingqian Li, Kissauo Tchedre, Christian Antolik, Jie Ma, Justin Chew, Tor Paaske Utheim, Xizhong A Huang, Honghua Yu, et al. Igfbpl1 regulates axon growth through igf-1-mediated signaling cascades. *Scientific reports*, 8(1):2054, 2018.

Hannah Hochgerner, Amit Zeisel, Peter Lönnerberg, and Sten Linnarsson. Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell RNA sequencing. *Nature Neuroscience*, 21:290–299, 2018. doi: 10.1038/s41593-017-0056-2.

Luni Hu, Qianqian Chen, Ping Qiu, Hua Qin, Yilin Zhang, Lei Cao, Tianyi Xia, Ziqing Deng, Yong Zhang, Shuangsang Fang, et al. Unicell: Towards a unified solution for cell annotation, nomenclature harmonization, atlas construction in single-cell transcriptomics. *bioRxiv*, 2025. doi: 10.1101/2025.05.06.652331.

Maria Eleni Kastriti, Louis Faure, Dorothea Von Ahsen, Thibault Gerald Bouderlique, Johan Boström, Tatiana Solovieva, Cameron Jackson, Marianne Bronner, Dies Meijer, Saida Hadjab, et al. Schwann cell precursors represent a neural crest-like state with biased multipotency. *The EMBO journal*, 41(17):e108780, 2022.

Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E Kastriti, Peter Lönnerberg, Alessandro Furlan, et al. Rna velocity of single cells. *Nature*, 560(7719):494–498, 2018.

Daniel Levine, Syed A Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil, David Zhang, Xingyu Chen, Sina Ghadermarzi, Ruiming Wu, Zihe Zheng, Ivan Vrkic, Anna Zhong, Daphne Raskin, Insu Han, Antonio Henrique De Oliveira Fonseca, Josue Ortega Caro, Amin Karbasi, Rahul Madhav Dhodapkar, and David Van Dijk. Cell2Sentence: Teaching large language models the language of biology. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 27299–27325. PMLR, 2024.

Mayu Nishikawa, Hideki Mori, and Masayuki Hara. Reduced zinc cytotoxicity following differentiation of neural stem/progenitor cells into neurons and glial cells is associated with upregulation of metallothioneins. *Environmental Toxicology and Pharmacology*, 39(3):1170–1176, 2015. doi: 10.1016/j.etap.2015.04.009.

Xiaojie Qiu, Yan Zhang, Jorge D Martin-Rufino, Chen Weng, Shayan Hosseinzadeh, Dian Yang, Angela N Pogson, Marco Y Hein, Kyung Hoi Joseph Min, Li Wang, et al. Mapping transcriptomic vector fields of single cells. *Cell*, 185(4):690–711, 2022.

Tijana Radic, Lara Frieß, Aruvi Vijikumar, Tassilo Jungenitz, Thomas Deller, and Stephan W Schwarzacher. Differential postnatal expression of neuronal maturation markers in the dentate gyrus of mice and rats. *Frontiers in neuroanatomy*, 11:104, 2017.

Natalí B Rasetto, Damiana Giacomini, Ariel A Berardino, Tomás Vega Waichman, Maximiliano S Beckel, Daniela J Di Bella, Juliana Brown, M Georgina Davies-Sala, Chiara Gerhardinger, Dieter Chichung Lie, et al. Transcriptional dynamics orchestrating the development and integration of neurons born in the adult hippocampus. *Science Advances*, 10(29):eadp6039, 2024.

Moritz Schaefer, Peter Peneder, Daniel Malzl, Salvo Danilo Lombardo, Mihaela Peycheva, Jake Burton, Anna Hakobyan, Varun Sharma, Thomas Krausgruber, Celine Sin, et al. Multimodal learning enables chat-based exploration of single-cell data. *Nature Biotechnology*, pp. 1–11, 2025.

Leonid Schneider. Survival of neural stem cells undergoing dna damage-induced astrocytic differentiation in self-renewal-promoting conditions in vitro. *PloS one*, 9(1):e87228, 2014.

Maciej Sypetkowski, Joanna Krawczyk, Łukasz Smoliński, Remigiusz Kinas, Przemysław Pietrzak, Tomasz Jetka, and Rafał Powalski. Omicslm: A multimodal large language model for multi-sample omics reasoning. *arXiv preprint arXiv:2605.06728*, 2026.

Weixu Wang, Zhiyuan Hu, Philipp Weiler, Sarah Mayes, Marius Lange, Daniel M Fountain, Julianna O Haug, Jingye Wang, Zhengyuan Xue, Tatjana Sauka-Spengler, et al. Regvelo: gene-regulatory-informed dynamics of single cells. *Cell*, 189(12):3773–3800, 2026.

Yafei Wang, Zhiheng Lu, Yilan Zhang, Yuqun Cai, Di Yun, Tianxiang Tang, Zheping Cai, Chunyang Wang, Yandong Zhang, Fang Fang, et al. Transcription factor 4 safeguards hippocampal dentate gyrus development by regulating neural progenitor migration. *Cerebral Cortex*, 30(5):3102–3115, 2020.

Zhijian Wei, Runze Ma, Zichen Wang, Zhongmin Li, Shuotong Song, and Shuangjia Zheng. VC-World: A biological world model for virtual cell simulation. In *International Conference on Learning Representations*, 2026.

Philipp Weiler, Marius Lange, Michal Klein, Dana Pe'er, and Fabian Theis. CellRank 2: unified fate mapping in multiview single-cell data. *Nature Methods*, 21(7):1196–1205, 2024. doi: 10.1038/s41592-024-02303-9.

Menghua Rachel Wu, Russell Littman, Jacob Levine, Lin Qiu, Tommaso Biancalani, David Richmond, and Jan-Christian Huetter. Contextualizing biological perturbation experiments through language. In *International Conference on Learning Representations*, volume 2025, pp. 73598–73624, 2025a.

Yicheng Wu, Vladyslav I Korobeynyk, Margherita Zamboni, Felix Waern, John Darby Cole, Sarah Mundt, Melanie Greter, Jonas Frisén, Enric Llorens-Bobadilla, and Sebastian Jessberger. Multimodal transcriptomics reveal neurogenic aging trajectories and age-related regional inflammation in the dentate gyrus. *Nature neuroscience*, 28(2):415–430, 2025b.

Chuxi Xiao, Yuang Ding, Haiyang Bian, Yixin Chen, Lei Wei, and Xuegong Zhang. Tokenizing single-cell transcriptomes as a native language for large language models. *bioRxiv*, 2025. doi: 10.1101/2025.10.22.684047.

Elliot Xie, Lingxin Cheng, Jack Shireman, Yujia Cai, Jihua Liu, Chitrasen Mohanty, Mahua Dey, and Christina Kendziorski. CASSIA: A multi-agent large language model for automated and interpretable cell annotation. *Nature Communications*, 17:389, 2026. doi: 10.1038/s41467-025-67084-x.

Jian Xue, Andrew T Brawner, Jacqueline R Thompson, Tushar D Yelhekar, Kyra T Newmaster, Qiang Qiu, Yonatan A Cooper, C Ron Yu, Yasir H Ahmed-Braima, Yongsoo Kim, et al. Spatiotemporal mapping and molecular basis of whole-brain circuit maturation. *bioRxiv*, 2024.

Chang Yu, Yuxuan Luo, Yixuan Du, Yuqing Zhou, Siyuan Li, Jingbo Zhou, Zhen Lei, Stan Z Li, et al. Geometric flow grounding: A unified manifold decoupling framework for dynamics discovery and verification. In *Forty-third International Conference on Machine Learning*, 2026.

Amit Zeisel, Hannah Hochgerner, Peter Lönnerberg, Anna Johnsson, Fatima Memic, Job Van Der Zwan, Martin Häring, Emelie Braun, Lars E Borm, Gioele La Manno, et al. Molecular architecture of the mouse nervous system. *Cell*, 174(4):999–1014, 2018.

Fan Zhang, Tianyu Liu, Zhihong Zhu, Hao Wu, Haixin Wang, Donghao Zhou, Yefeng Zheng, Kun Wang, Xian Wu, and Pheng-Ann Heng. CellVerse: Do large language models really understand cell biology? In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2272. Datasets and Benchmarks Track.

Juan Zhang, Fen Ji, Yanli Liu, Xuepei Lei, Hong Li, Guangju Ji, Zengqiang Yuan, and Jianwei Jiao. Ezh2 regulates adult hippocampal neurogenesis and memory. *The Journal of neuroscience*, 34 (15):5184–5199, 2014.

A Appendix

A.1 TransBench Construction

This subsection documents dataset harmonization and sample construction omitted from the concise TransBench description in Section 4.1.

All task labels are constructed independently of the evaluated velocity values. The raw expression matrix and cell annotations determine the local cell graph and the source–target pairs. The processed RegVelo, DeepVelo, and veloVI objects are consulted only to identify genes for which the corresponding baseline provides a velocity value. The task builder never reads the predicted velocity entries, transition matrices, CBDir scores, or latent time when constructing reference labels.

For each dataset, we construct a 50-NN graph from a 30-dimensional uncentered truncated-SVD representation of the raw expression matrix, excluding the query cell itself and using seed 42. Hematopoiesis and MouseBrain use their provided log-transformed expression. DentateGyrus contains counts, so we apply library-size normalization to 10^4 followed by $\log(1+x)$ before graph construction and expression differences. Cluster labels are canonicalized before matching the dataset-specific directed reference edges.

TE and HTE use the same examples, containing 200 forward-reference transitions and 100 explicitly reversed transitions. Unlisted cluster pairs are not silently treated as negatives. NSP candidates are the distinct cluster identities among the source cell's raw-expression neighbors. For each candidate, the prompt shows genes ranked by mean expression among those local neighbor cells rather than an artificial global cluster centroid. A source is rejected if fewer than two candidate states are present or if its neighborhood contains more than one known positive destination. This strict filter yields 145

Table 5: Detailed TransBench task-construction parameters. "Delta" denotes target-minus- source expression on the declared task scale.

Track	Reference unit	Selection rule	Model-visible evidence	Output
TE	directed cell pair	200 forward + 100 reverse pairs	clusters; two ranked top-gene lists	yes/no
HTE	same pairs as TE	cluster names hidden; all clusters are candidates	two ranked top-gene lists; candidate clusters	clusters + yes/no
NSP	source boundary cell	≥ 2 local candidate states; ≤ 1 known future state	source genes; neighbor-cluster mean-gene lists	one candidate state
GVD	pair–gene item	$\lvert\Delta\rvert \geq 0.5$; relative change ≥ 0.25; expression ≥ 1; at most 3 genes per pair	clusters; masked gene; two ranked top-gene lists	up/down
IGR	directed cell pair	100 shuffled candidates; top 20 positive deltas	clusters; masked context; candidate genes	ordered 20 genes
DGR	same pair and pool as IGR	100 shuffled candidates; top 20 negative deltas	clusters; masked context; candidate genes	ordered 20 genes

Hematopoiesis and 215 MouseBrain NSP examples; we do not fill these shortfalls with ambiguous samples.

For GVD, the sign of the local expression difference defines the reference class; 150 increasing and 150 decreasing examples are selected per dataset. The queried gene is removed from both visible top-gene lists. IGR and DGR use the same 300 cell pairs and the same deterministically shuffled 100-gene candidate pool. Each pool contains the strongest increasing and decreasing genes and difficult distractors selected near the ranking boundary, from high-expression genes, and from genes with large absolute changes. Every candidate gene is removed from the visible source and target top-gene context. The reference ranking contains the 20 strongest changes in the requested direction. The eligible gene universes contain 109, 692, and 575 genes for Hematopoiesis, MouseBrain, and DentateGyrus, respectively.

A.2 TransBench Prompt Specification

This subsection reports the exact model-visible prompt structure and the P1–P4 ablation components. Templates are shown instead of individual cell records; the released JSONL files preserve every complete prompt verbatim.

A.2.1 Shared Prompt Components

The following framed templates reproduce the model-visible structure. Square- bracketed terms are placeholders. Ellipses abbreviate long arrays only for the paper; the actual prompt contains every listed gene.

Common system message (all profiles)

```
You are analyzing cell-state transitions in [species] [biological
system].

Use only the supplied task input, your general biological
knowledge, and any external biological knowledge explicitly
included in the user message.  Do not invent measurements that are
not provided.

For this task, ``top genes'' means genes with the highest observed
expression, not highly variable genes.  Every top-gene list is
ordered from highest to lowest observed expression within the
provided gene universe.  The absence of a gene from a top-gene list
does not imply zero expression.

[Reasoning guidance is inserted here only for P3 and P4.]

Return exactly one valid JSON object.  Do not include markdown,
explanations, analysis, or any text outside the JSON object.
```

Common user-message JSON skeleton

```
{
  ``task_id'': ``[T1/T1-H/T2/T4/T5/T6]'',
  ``task_name'': ``[task name]'',
  ``instruction'': ``[task-specific instruction]'',
  ``input'': {[task-specific visible fields]},
  ``required_output_schema'': {[schema]},
  ``rules'': [``one JSON object'', ``exact candidate spelling'',
``confidence in 0--1''],
  [``allowed_labels'' or ``allowed_candidates'' or
``ranking_rules''],
  [``external_biological_knowledge'': {...}] only in P2/P4,
  [``reasoning_guidance'': {...}] only in P3/P4
}
```

The common input prefix records the species, biological system, definition and ordering of top genes, gene-universe size, and the warning that absence from a top-gene list does not mean zero expression. Expression magnitudes are never shown.

A.2.2 Task-Specific Prompt Components

TE — Transition Existence

Instruction: decide whether a direct directed transition from the named source cluster to the named target cluster is biologically supported. The cells are KNN neighbors, but adjacency alone does not establish direction or ancestry.

Input: {``source_cluster'': [state], ``target_cluster'': [state], ``source_top_genes'': $[g_1^s, g_2^s, \ldots]$, ``target_top_genes'': $[g_1^t, g_2^t, \ldots]$}

Allowed labels: [``yes'', ``no'']

Output: {``label'': ``yes|no'', ``confidence'': [0--1]}

HTE — Hidden-state Transition Existence

Instruction: infer the hidden source and target identities from the ordered single-cell gene lists and candidate cluster names, then decide whether the directed transition is supported. The first list is always the source cell.

Input: {``source_top_genes'': $[g_1^s, \ldots]$, ``target_top_genes'': $[g_1^t, \ldots]$, ``candidate_clusters'': $[c_1, \ldots, c_m]$}

Output: {``source_cluster'': $c_i$, ``target_cluster'': c_j, ``label'': ``yes|no'', ``confidence'': [0--1]}

Scoring note: only label contributes to HTE accuracy and Macro-F1; the two inferred cluster fields are retained for diagnosis but are unscored.

NSP — Next-State Prediction

Instruction: select the most plausible immediate next-state cluster for the source boundary cell. Candidate features are ranked mean-expression genes from the source cell's local KNN neighbors grouped by candidate cluster.

Input: {''current_cluster'': [state], ''current_top_genes'': [g_1,...], ''candidate_clusters'': [c_1,...,c_m], ''candidate_cluster_features'': {c_1: [$g_1^{c_1}$,...], ..., c_m: [$g_1^{c_m}$,...]}}

Output: {''label'': ''one exact candidate cluster'', ''confidence'': [0--1]}

GVD — Gene Velocity Direction

Instruction: predict whether the specified gene increases or decreases from the source cell toward its target-neighbor cell. The queried gene is intentionally removed from both visible top-gene lists; stable is not an allowed class.

Input: {''source_cluster'': [state], ''target_cluster'': [state], ''gene'': [masked query gene], ''source_top_genes'': [g_1^s,...], ''target_top_genes'': [g_1^t,...]}

Output: {''label'': ''up|down'', ''confidence'': [0--1]}

IGR — Increasing-Gene Ranking

Instruction: rank the unordered candidate genes from strongest to weakest predicted increase. All candidates are removed from both visible top-gene lists.

Input: {''source_cluster'': [state], ''target_cluster'': [state], ''source_top_genes'': [g_1^s,...], ''target_top_genes'': [g_1^t,...], ''candidate_genes'': [q_1,...,q_{100}], ''candidate_pool_order'': ''unordered'', ''top_k'': 20}

Output: {''genes'': [exactly 20 distinct candidates in predicted order], ''confidence'': [0--1]}

DGR — Decreasing-Gene Ranking

Instruction and input are identical to IGR except that genes are ordered from the strongest to the weakest predicted decrease.

Output: {''genes'': [exactly 20 distinct candidates in predicted order], ''confidence'': [0--1]}

For IGR and DGR, masking can leave an empty source or target top-gene list in a sparse cell. The prompt explicitly states that this does not imply absence of expression. The 100 candidate genes are complete in the actual input and their order is unrelated to the reference ranking.

A.2.3 RAG Component in P2 and P4

The RAG block contains two levels. Dataset-shared knowledge is identical for every P2/P4 example from a dataset: 50 highly variable genes, a bounded TF–target prior, and five task-safe literature summaries. Task-specific Cell Ontology records are then selected from model-visible clusters. TE, GVD, IGR, and DGR use the visible source and target states; NSP uses the source and all candidates; HTE uses all candidate states so that ontology selection cannot reveal the hidden answer.

RAG addition to the user JSON (P2 and P4 only)

```
''external_biological_knowledge'': {
  ''shared_knowledge'': {
    ''hvg_context'': {''definition'': [text], ''hvg_count'': 50,
''genes'': [h1,...,h50]},
    ''tf_target_prior'': {''database'': ''TRRUST'', ''edge_count'':
[count], ''effect_counts'': {...}, ''interpretation'':
[non-deterministic prior warning], ''edges'': [{''tf'': [TF],
''target'': [gene], ''effect'': [effect]},...]},
    ''task_safe_literature_summaries'': [[summary 1], ..., [summary
5]]
  },
  ''task_relevant_cell_ontology'': [{
    ''task_cluster'': [cluster], ''ontology_name'': [name],
''ontology_id'': [CL identifier], ''definition'': [definition],
''hierarchy'': [[root], ..., [mapped cell type]],
''mapping_relation'': [exact/broad/composite]
  },...]
}
```

The evaluated Hematopoiesis prompts contain 83 TF–target records: 35 activating, 19 inhibiting, and 29 with unspecified effect. The evaluated MouseBrain prompts contain 17 retained directed edges (12 activating and 5 inhibiting), and the DentateGyrus prompts contain 34 (27 activating and 7 inhibiting). For the two neural datasets, unknown-effect edges, TFs absent from the measured gene universe, and regulator–target pairs with conflicting directions are removed. We report the Hematopoiesis composition explicitly because its already-evaluated prompt version retained unknown-effect records; they are presented as non-deterministic literature associations rather than directional evidence.

The five summaries are shared across all tasks within a dataset. They describe cell identities, marker programs, regulatory mechanisms, and developmental context, but exclude TransBench sampled cell pairs, reference labels, candidate order, numerical expression differences, velocity predictions, and TransBench performance. The RAG instruction states that external knowledge is supporting context and not a deterministic answer key.

A.2.4 Reasoning Component in P3 and P4

Reasoning additions (P3 and P4 only)

```
System-message addition:
Reason internally and concisely, using at most 2560 tokens as a
soft reasoning budget.  Do not restate the full input or discuss
every candidate.  Stop once the evidence is sufficient, then return
the complete required JSON object.

User-JSON addition:
''reasoning_guidance'': {
  ''mode'': ''internal_concise_reasoning'',
  ''reasoning_soft_limit'': 2560,
  ''requirements'': [''do not restate the context'', ''stop when
evidence is sufficient'', ''return only JSON'']
}
```

The 2560-token value is a soft textual constraint, not a guaranteed decoder limit. At the API level, P1/P2 explicitly disable thinking. P3/P4 enable thinking with a budget of 2048 for Qwen3.8-Flash and Alibaba-direct Kimi-K2.6. For DeepSeek-V4.1-Flash, the gateway sends `reasoning_effort=low` and explicitly allowlists that parameter. We do not set `max_tokens`

Table 6: Composition of the four TransBench prompt profiles. The task template is part of the base component and is identical across profiles for a given example.

Profile	Base system + task input	External RAG	Reasoning prompt	API thinking
P1	Yes	No	No	Disabled
P2	Yes	Yes	No	Disabled
P3	Yes	No	Yes	Enabled
P4	Yes	Yes	Yes	Enabled

or `max_completion_tokens`, because a hard total-output cap could terminate reasoning before the required JSON is emitted.

A.2.5 Modular Prompt Composition

The preceding boxes display the shared, task-specific, RAG, and reasoning components separately. This modular presentation avoids repeating nearly identical complete prompts for P1–P4 and makes the controlled ablations explicit. It also keeps the exact model-visible template auditable while separating fixed input content from the two experimental interventions.

For a task-specific base prompt B_t, external knowledge block K, and reasoning block R, the four conditions are $P1 = B_t$, $P2 = B_t + K$, $P3 = B_t + R$, and $P4 = B_t + K + R$. The model receives one system message and one user message. Gold labels and evaluation metadata remain separate from both.

A.3 TransBench Evaluation Protocol and Reproducibility

This subsection describes the safeguards, parsing rules, metrics, matched RNA-velocity baselines, and artifacts used to reproduce the evaluation.

A.3.1 Knowledge-Leakage Prevention

Prompt construction follows a positive whitelist rather than deleting fields from the complete task record. Only fields displayed in the prompt components above can enter the two model messages. The builder rejects any model-visible occurrence of gold labels, metadata, expression signatures or magnitudes, source/target cell identifiers, pair identifiers, expression differences, transition matrices, velocity scores, and other evaluation-only fields. It also verifies that the RAG block appears only in P2/P4 and that the reasoning block appears only in P3/P4.

Task-specific masking prevents direct answer recovery. The GVD query gene is absent from both top-gene lists. Every IGR/DGR candidate is absent from both lists, candidate genes are unique, and their order is deterministically shuffled. The external knowledge is fixed at the dataset level except for ontology records selected exclusively from visible state names. No task label, expression delta, velocity, model prediction, or metric is used for retrieval. Prompt files, knowledge blocks, ontology resources, source task files, and generated outputs are hashed to make post hoc changes detectable.

A.3.2 Structured Prediction, Parsing, and Failure Handling

Every prompt requests exactly one JSON object. TE and GVD require labels from fixed two-class vocabularies; NSP requires an exact candidate-cluster string; IGR/DGR require exactly 20 distinct genes drawn from the supplied pool. The runner first attempts strict JSON parsing, optionally removes one surrounding Markdown JSON fence, and otherwise extracts the first syntactically valid JSON object. Schema and candidate constraints are validated after parsing. Invalid outputs are not silently converted into correct answers.

Requests use JSON response mode when supported. Transient gateway failures are retried with a finite retry count and capped exponential backoff; non-retryable errors and exhausted retries are written to `errors.jsonl`. A failed example does not terminate the remaining run. Resumption is keyed by task ID, model ID, and a hash of the complete message list, preventing stale predictions

from being reused after a prompt change. Coverage is reported for every profile–task pair, and partial evaluations are identified in the result table.

A.3.3 METRICS AND MATCHED VELOCITY BASELINES

TE, HTE, NSP, and GVD are evaluated using accuracy and Macro-F1, with Macro-F1 computed over classes that occur in the reference labels for that evaluated task. For HTE, only the yes/no label is scored. IGR and DGR use NDCG@20 as the primary metric. Overlap@20, candidate-violation rate, and duplicate-gene rate are retained as diagnostics but are not used as primary performance measures.

The matched RegVelo, DeepVelo, and veloVI baselines use existing per-cell velocity outputs and are not rerun during evaluation. For NSP, candidate cells are reconstructed from the same raw-expression KNN recipe. For each individual candidate cell j, the baseline computes

$$s_{ij} = \cos(\mathbf{v}_i, \mathbf{x}_j - \mathbf{x}_i)\,,$$

and predicts the cluster of the highest-scoring candidate cell; it does not replace candidate cells with a cluster-average expression vector. For GVD, the sign of the queried component of $\mathbf{v}_i$ gives the predicted direction. For IGR and DGR, the supplied candidate genes are sorted by the corresponding components of $\mathbf{v}_i$ in descending and ascending order, respectively. Velocity baselines are not defined for TE or HTE.

A.3.4 TRANSBENCH REPRODUCIBILITY CHECKLIST

For a complete release, we preserve: (i) raw and processed H5AD checksums; (ii) cluster keys, gene-identifier columns, preprocessing scale, and normalization; (iii) reference and reverse edges; (iv) KNN/SVD parameters and seed; (v) task JSONL files and build reports; (vi) RAG sources, filtering counts, ontology mappings, and literature summaries; (vii) every P1–P4 system/user message and its SHA-256 hash; (viii) exact gateway model identifiers and request controls; (ix) raw parsed predictions, usage metadata, and error logs; and (x) evaluation code, per-task coverage, class universes, and metric outputs. These artifacts distinguish a missing prediction from an incorrect one and make the reported ablations reproducible without exposing evaluation labels to the model.

A.4 VELOCITY REFINER AND TRAINING OBJECTIVES

Neural network architecture. We employ a multi-layer perceptron (MLP) as the velocity refiner f_θ. Its architecture consists of several fully connected layers, interleaved with LayerNorm, GELU activation, and dropout (rate 0.1). The network takes the raw expression vector $\mathbf{x}_i$ as input and outputs a refined velocity $\hat{\mathbf{v}}_i$.

Training objectives. The training objective combines three components:

$$\mathcal{L} = \lambda_{\text{mse}}\mathcal{L}_{\text{mse}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}} + \lambda_{\text{edge}}\mathcal{L}_{\text{edge}}. \tag{1}$$

Base-velocity reconstruction loss $\mathcal{L}_{\text{mse}}$ anchors the predicted velocities to the base estimator's cell-wise output:

$$\mathcal{L}_{\text{mse}} = \frac{1}{NG}\sum_{i=1}^{N}\sum_{g=1}^{G}\left(\hat{v}_{ig} - v_{ig}^{(0)}\right)^2. \tag{2}$$

Within-state smoothness loss $\mathcal{L}_{\text{smooth}}$ averages cosine dissimilarity over connected cells sharing the same state label. With $\mathcal{A}_{\text{same}} = \{(i,j) : A_{ij} > 0,\ c_i = c_j\}$, it is

$$\mathcal{L}_{\text{smooth}} = \frac{1}{|\mathcal{A}_{\text{same}}|}\sum_{(i,j)\in\mathcal{A}_{\text{same}}}\left(1 - \mathrm{n}(\hat{\mathbf{v}}_i)^\top \mathrm{n}(\hat{\mathbf{v}}_j)\right), \tag{3}$$

where A is the cell-connectivity matrix and n denotes the normalization defined in Table 2.

Edge-prior loss $\mathcal{L}_{\text{edge}}$ is specified in Table 2.

A.5 LOCAL ODE FIDELITY DIAGNOSTIC

From RNA kinetics to local compatibility. RNA velocity describes the derivative of spliced RNA abundance under a transcription–splicing–degradation model (La Manno et al., 2018; Bergen et al., 2020). To accommodate changes in kinetics along a developmental trajectory, we allow the rates to depend on latent developmental time t, represented by a pseudotime coordinate:

$$\frac{du_g(t)}{dt} = \alpha_g(t) - \beta_g(t)u_g(t), \qquad v_g(t) = \frac{ds_g(t)}{dt} = \beta_g(t)u_g(t) - \gamma_g(t)s_g(t), \tag{4}$$

where $\alpha_g(t)$, $\beta_g(t)$, and $\gamma_g(t)$ are effective transcription, splicing, and degradation rates in this time coordinate. Inferring these rate functions and trajectories from snapshot observations is underdetermined: different kinetic parameterizations and latent-time assignments can explain the same expression data. Globally constant gene-specific rates form a restricted special case. For example, the steady-state mode implemented in scVelo estimates a shared steady-state ratio for each gene, whereas its dynamical mode relaxes the steady-state assumption while retaining gene-specific kinetic parameters (Bergen et al., 2020). Constant rates and steady-state expression are therefore distinct assumptions.

Rather than reconstructing a complete trajectory, we assess whether an already predicted velocity is locally compatible with the spliced-RNA equation. Since the datasets do not directly supply each cell's latent developmental time, we use neighboring expression states as a local approximation. For cell i, let $\mathcal{N}_i$ contain up to 30 graph neighbors, excluding i. We assume that the splicing and degradation rates vary sufficiently slowly to be approximately constant within this neighborhood:

$$\beta_g(t_j) \approx \beta_{ig}, \qquad \gamma_g(t_j) \approx \gamma_{ig}, \qquad j \in \mathcal{N}_i. \tag{5}$$

Under this approximation, compatible velocities satisfy $\hat{v}_{jg} \approx \beta_{ig}u_{jg} - \gamma_{ig}s_{jg}$. Thus, the local observations $(u_{jg}, s_{jg}, \hat{v}_{jg})$ should lie near a plane. Checking this relation requires a local linear fit rather than integration of the ODE or estimation of a global pseudotime trajectory.

Derivation of the projection score. The diagnostic fits the affine relation $\hat{v}_{jg} = a_{ig} + b_{u,ig}u_{jg} + b_{s,ig}s_{jg} + \varepsilon_{jg}$, allowing a local offset in the observed quantities. In the kinetic interpretation, $b_{u,ig}$ corresponds to β_{ig} and $b_{s,ig}$ to $-\gamma_{ig}$. After removing non-finite observations, we subtract neighborhood means and define

$$\mathbf{X}_{ig} = \left[(u_{jg} - \bar{u}_{ig},\ s_{jg} - \bar{s}_{ig})\right]_{j\in\mathcal{N}_i}, \qquad \mathbf{y}_{ig} = \left[\hat{v}_{jg} - \bar{v}_{ig}\right]_{j\in\mathcal{N}_i}. \tag{6}$$

Centering eliminates the intercept, giving $\mathbf{y}_{ig} \approx \mathbf{X}_{ig}\mathbf{b}_{ig}$. The least-squares coefficients and fitted velocity are

$$\mathbf{b}^{\star}_{ig} = \mathbf{X}^{+}_{ig}\mathbf{y}_{ig}, \qquad \widehat{\mathbf{y}}_{ig} = \mathbf{X}_{ig}\mathbf{X}^{+}_{ig}\mathbf{y}_{ig}, \tag{7}$$

where $\mathbf{X}^{+}_{ig}$ is the pseudoinverse. The residual $\mathbf{r}_{ig} = \mathbf{y}_{ig} - \widehat{\mathbf{y}}_{ig}$ is orthogonal to the fitted vector, so $\|\mathbf{y}_{ig}\|_2^2 = \|\widehat{\mathbf{y}}_{ig}\|_2^2 + \|\mathbf{r}_{ig}\|_2^2$. The explained fraction is consequently equivalent to the usual centered coefficient of determination, $1 - \|\mathbf{r}_{ig}\|_2^2/\|\mathbf{y}_{ig}\|_2^2$. With numerical stabilization, the implemented score is

$$R^2_{ig} = \mathrm{clip}_{[0,1]}\left(\frac{\|\mathbf{X}_{ig}\mathbf{X}^{+}_{ig}\mathbf{y}_{ig}\|_2^2}{\|\mathbf{y}_{ig}\|_2^2 + 10^{-8}}\right), \tag{8}$$

Higher scores indicate that local variation in the predicted velocity is better explained by the observed unspliced and spliced abundances. We require at least five valid neighbors and nonzero velocity variance. When the condition number of $\mathbf{X}_{ig}$ exceeds 50, the implementation projects onto its leading singular direction instead of using both columns; undefined scores are excluded.

Interpretation and visual inspection. The fit does not impose $\beta_{ig} \geq 0$ or $\gamma_{ig} \geq 0$, equivalently $b_{u,ig} \geq 0$ and $b_{s,ig} \leq 0$. The score therefore measures local linear compatibility rather than establishing physically admissible kinetic rates. It is also sensitive to noise in the supplied u/s measurements, which affects the fitted plane and its residuals. A global mean can obscure this spatial heterogeneity and is insufficient on its own to assess kinetic fidelity. We therefore emphasize visual inspection of local score patterns and residual geometry in Figure 4. The UMAP colors show R^2_{ig} for a representative gene selected once and shared across both runs, from genes with valid scores in at least half the cells. Selection favors both dispersion (90th minus 10th percentile) and entropy across the score distribution. These maps reveal where local compatibility is retained or lost.

A.6 Effect of Prior Source on Velocity Refinement

We compare the literature-curated transition prior with three language-model outputs on Dentate Gyrus using scVelo as the base velocity model. Each output replaces the curated edge list while the refinement method remains the same. Table 7 lists the supplied transitions and the resulting velocity metrics.

Table 7: Effect of prior source on veloLLM. The model-output column lists the directed cell-state transitions supplied as the edge prior. The literature-curated prior was assembled manually from the dataset's associated literature. Higher is better for all three metrics; bold indicates the highest value in each column.

Prior source	Model output (directed transitions)	VeloCoh	ICCoh	CBdir
Dentate Gyrus				
Literature-curated	nIPC → Neuroblast; Neuroblast → Granule immature; Granule immature → Granule mature; Radial Glia-like → Astrocytes	0.7627	0.8286	**0.7285**
qwen3.8-flash	Neuroblast → Granule immature; Granule immature → Granule mature; Neuroblast → Mossy; Radial Glia-like → Astrocytes	0.7716	**0.8316**	0.4029
deepseek-v4.1-flash	Radial Glia-like → Astrocytes; Neuroblast → Granule immature; Granule immature → Granule mature	0.7670	0.8283	0.4082
glm-5.3 & kimi-k2.6	Neuroblast → Granule immature; Granule immature → Granule mature	**0.7756**	0.8268	0.0252

The language-model priors yield VeloCoh values of 0.7670–0.7756 and ICCoh values of 0.8268–0.8316, close to the literature-curated values of 0.7627 and 0.8286. Their CBdir scores are 0.4029, 0.4082, and 0.0252, compared with 0.7285 for the literature-curated prior. All three model outputs omit the nIPC → Neuroblast transition included in the curated graph, and the graphs differ in other edges as shown in the table. In this Dentate Gyrus comparison, directional performance is sensitive to the supplied transition prior.

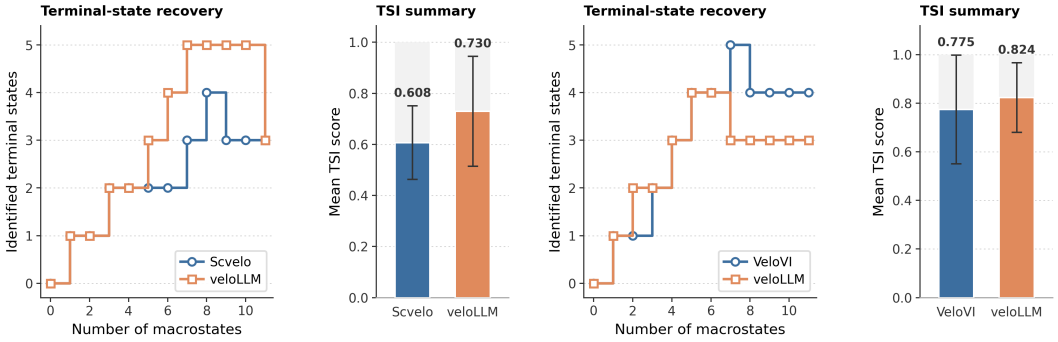
A.7 Terminal-State Identification Details and Additional Results

We use a transition kernel that combines velocity and connectivity kernels with weights of 0.8 and 0.2. For each stability threshold t_j and macrostate count m, GPCCA predicts terminal states; we collapse split macrostate names to their parent cell-type labels and count the distinct predictions matching the five annotated terminal cell types. The left-step area under this recovery curve is divided by the area under an ideal curve that recovers one additional terminal type per macrostate until all five are found. The TSI score is the mean of these normalized areas across 20 thresholds, and the error bars show their standard deviation. Panel 7d gives the calculation used by the evaluation script. If GPCCA fails at a macrostate count, the preceding recovery count is carried forward (or zero at the first count).

Figure 7 extends the Hematopoiesis comparison to scVelo, veloVI, and RegVelo. Mean TSI increases from 0.608 to 0.730 with scVelo and from 0.775 to 0.824 with veloVI, whereas the RegVelo comparison remains similar (0.846 versus 0.840). Thus, the directional improvements in Table 3 do not imply a TSI gain for every base estimator or at every macrostate resolution.

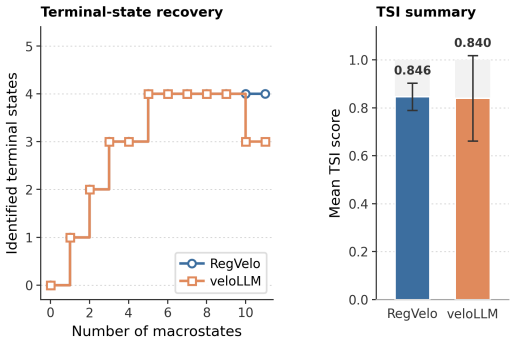
Table 8: Ablation study on the hidden layer configurations.

Hidden layer	Params	FLOPs	VeloCoh	ICCoh	CBdir
(256, 256)	456.439K	456.192K	0.8881	0.7053	0.8345
(512, 512)	1.043M	1.043M	0.8756	0.6646	0.8357
(256, 512, 256)	654.583K	654.848K	0.8830	0.6903	0.8334
(256, 512, 512, 256)	918.263K	919.040K	0.8803	0.6862	0.8342



(a) scVelo

(b) veloVI



(c) RegVelo

Quantity	Computation
Kernel	$P = 0.8P_{\text{vel}} + 0.2P_{\text{conn}}$
Threshold	$t_j = 0.1 + 0.045j,\ j = 0, \ldots, 19$
Recovery	$r_j(m) = \lvert\widehat{\mathcal{T}}_j(m) \cap \mathcal{T}^\star\rvert$
Ideal area	$A^\star = \sum_{m=0}^{10} \min(m, 5) = 40$
Area ratio	$q_j = \frac{1}{40} \sum_{m=0}^{10} r_j(m)$
TSI	$\frac{1}{20} \sum_{j=0}^{19} q_j$
Error bar	$\mathrm{SD}(q_0, \ldots, q_{19})$

(d) TSI score computation

Figure 7: Additional terminal-state identification comparisons on Hematopoiesis. Panels (a)–(c) show recovery curves and mean TSI scores for three base velocity models; panel (d) gives the score computation, where $m = 1, \ldots, 11$, $r_j(0) = 0$, and $\mathcal{T}^\star$ is the set of five annotated terminal cell types.

Scaling Analysis. We further investigate how the model capacity affects the performance and computational cost of veloLLM by varying the configuration of its hidden layers. Specifically, we consider four configurations, including $(256, 256)$, $(512, 512)$, $(256, 512, 256)$, and $(256, 512, 512, 256)$. As shown in Table 8, increasing the hidden layer size leads to a corresponding increase in the number of trainable parameters and computational cost, measured by Params and FLOPs, respectively. We compare these configurations in terms of VeloCoh, ICCoh, and CBdir to characterize the trade-off between model capacity and velocity inference performance.

The results show that veloLLM maintains competitive performance across different hidden layer configurations, while larger configurations introduce additional computational overhead. These results provide a practical reference for selecting the model size according to the desired balance between predictive performance and computational efficiency.

A.8 Perturbation Sensitivity and Biological Associations

For each candidate gene g, we replace its expression in each cell with its mean expression in the source cell type of the corresponding transition, leaving all other gene expression values unchanged, and re-run the model. Let $\mathbf{v}_i$ and $\mathbf{v}_i^{(-g)}$ denote the original and perturbed velocity vectors for cell i. The perturbation-induced angular deviation is

$$D_{i,g} = \arccos\left(\frac{\mathbf{v}_i^\top \mathbf{v}_i^{(-g)}}{\|\mathbf{v}_i\|_2 \|\mathbf{v}_i^{(-g)}\|_2}\right). \tag{9}$$

For transition $A \to B$, the gene-level sensitivity score averages this deviation over the $N_{A\to B}$ cells associated with the transition:

$$D_g^{A\to B} = \frac{1}{N_{A\to B}} \sum_{i \in A\to B} D_{i,g}. \tag{10}$$

The score is measured in radians. A larger value indicates greater sensitivity of the predicted velocity direction to perturbation of gene g and prioritizes it for investigation in the corresponding cell-state transition.

Biological associations of the prioritized genes. The prioritized genes in Figure 6 show biologically coherent, but varying-strength, associations with the cellular transitions considered here. The strongest correspondence is observed for *Igfbpl1*, which ranks first for the nIPC → neuroblast transition and second for the neuroblast → immature granule-cell transition. Multiple independent transcriptomic studies place *Igfbpl1* specifically within this early neuronal differentiation continuum. Hochgerner et al. identified two sequential neuroblast populations immediately downstream of proliferating nIPCs and found *Igfbpl1* to be expressed in both; importantly, they reported that *Igfbpl1* delineated these neuroblast states more specifically than the commonly used markers *Dcx* and *Sox11* (Hochgerner et al., 2018). Consistent with this observation, the broader nervous-system atlas of Zeisel et al. identified *Igfbpl1*-expressing neuroblast populations in the dentate gyrus and other neurogenic regions and concluded that *Igfbpl1* spans relatively early and later neuroblast states, rather than marking only a single discrete population (Zeisel et al., 2018). More recent analyses further reinforce this association: Wu et al. explicitly used *Igfbpl1* as a neuroblast-specific gene when resolving adult dentate-gyrus neurogenic lineages (Wu et al., 2025b), while temporally resolved profiling of adult-born granule cells found *Igfbpl1* already present during neuronal commitment in proliferative progenitors, reaching maximal expression in the first postmitotic neuroblast states and remaining detectable in subsequent immature granule-cell populations (Rasetto et al., 2024). Thus, its high ranking at both early transitions is consistent with a gene whose expression accompanies cell-cycle exit, neuroblast identity, and early neuronal maturation. Functional evidence outside the dentate-gyrus lineage also connects IGFBPL1 to neuronal morphogenesis: perturbation and supplementation experiments have shown that IGFBPL1 regulates neurite and axonal outgrowth, including in cultured hippocampal neurons (Guo et al., 2018). This functional observation provides additional biological plausibility, although it does not demonstrate that IGFBPL1 itself causally controls the nIPC-to-neuroblast or neuroblast-to-immature-granule-cell transitions in vivo.

The ranking of *Calb1* at the immature → mature granule-cell transition is similarly well supported as a marker of neuronal maturation. In the postnatal dentate gyrus, Radic et al. contrasted DCX as an immature-neuron marker with calbindin/Calb1 as a mature-neuron marker and observed a progressive shift from DCX-positive to calbindin-positive granule cells during granule-cell maturation (Radic et al., 2017). Transcriptomic profiling of adult-born granule cells likewise showed reduced immature-neuronal signatures followed by increased *Calb1* expression in young granule cells approaching mature states (Rasetto et al., 2024). In addition, whole-brain circuit-maturation experiments use Calb1 as a granule-cell maturation marker in the dentate gyrus (Xue et al., 2024). The high model sensitivity to *Calb1* at this transition therefore agrees with a reproducible change in granule-cell maturation state. These studies primarily establish *Calb1* as a readout of maturity, however, rather than as a demonstrated driver of the maturation process.

The early-transition rankings of *Tcf4* and *Ezh2* have more indirect but functionally relevant support. *Tcf4* is highly expressed in dentate neural progenitors during development, and conditional loss of *Tcf4* disrupts progenitor migration and dentate-gyrus formation, in part through regulation of *Wnt7b* (Wang et al., 2020). This establishes a role for *Tcf4* in dentate progenitor behavior, although

the experiments address developmental progenitor migration rather than specifically the nIPC → neuroblast transition represented in our analysis. *Ezh2* is more directly associated with proliferative intermediate progenitors: most Tbr2-positive transit-amplifying progenitor cells express Ezh2, Ezh2 is enriched in actively dividing NSCs/progenitors rather than quiescent NSCs, and its deletion reduces progenitor proliferation and subsequent neuronal production (Zhang et al., 2014). The same study further reported stronger Ezh2 immunoreactivity in proliferating $DCX^{+}Ki67^{+}$ neuroblasts than in $DCX^{+}Ki67^{-}$ immature neurons. Together, these findings make the early rankings of *Tcf4* and particularly *Ezh2* compatible with known progenitor programs, while stopping short of assigning either gene a transition-specific causal role.

Several additional prioritized genes have more limited or context-dependent literature support and are therefore best interpreted cautiously. *Stmn2* is associated with neuronal specification: in the adult dentate-gyrus trajectory of Wu et al., neuronal-specification genes including *Stmn2* distinguish activated/progressing neurogenic cells from quiescent NSCs, and *Stmn2* expression increases along the neurogenic activation trajectory (Wu et al., 2025b). This supports its association with commitment toward a neuronal state, but does not by itself establish STMN2 as a fate-determining regulator. Evidence for *Atp1a2* is less transition-specific. Although *Atp1a2* is strongly associated with astroglial transcriptional programs, it is also expressed in radial glia-like and dormant neural stem-cell states, making it more appropriate to interpret its prioritization as reflecting a shared glial-like molecular identity than as specific evidence for astrocytic differentiation (Hochgerner et al., 2018). Evidence for *Bcl2* and *Mt1* is weaker still and comes from experimental contexts outside normal dentate-gyrus lineage progression. BCL2 overexpression was reported to suppress apoptosis and thereby increase the efficiency of astroglial differentiation in irradiated neural stem-cell cultures (Schneider, 2014); this supports an association with survival during experimentally induced astrocytic differentiation, rather than a direct physiological role in specifying astrocyte fate. Similarly, differentiation of cultured neural stem/progenitor cells into neuronal and glial progeny increased expression of metallothionein-family genes including *Mt1* (Nishikawa et al., 2015). Given the broad stress-response functions of metallothioneins and the mixed differentiated population in that experiment, *Mt1* is better viewed as a general differentiation- and stress-associated program than as a marker of a particular lineage transition. Overall, the literature support is therefore strongest where the model prioritizes established stage-associated genes such as *Igfbpl1* and *Calb1*, whereas the remaining genes provide progressively more indirect biological concordance. These associations support the biological interpretability of the model sensitivities, but they should not be taken as evidence that each prioritized gene causally regulates the corresponding transition.