GraphVelo allows for accurate inference of multimodal velocities and molecular mechanisms for single cells

Yuhao Chen[1,2,a], Yan Zhang[2,a], Jiaqi Gan[2,a], Ke Ni[2], , Ming Chen[1,3,b], Ivet Bahar[4,b], Jianhua Xing[2,5,6,b]

1. *Department of Bioinformatics, College of Life Sciences, Zhejiang University, Hangzhou, 310058, China*
2. *Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA*
3. *Zhejiang Key Laboratory of Multi-omics Precision Diagnosis and Treatment of Liver Diseases, Department of General Surgery, Sir Run-Run Shaw Hospital, Zhejiang University School of Medicine, 310016, Hangzhou, China.*
4. *Laufer Center for Physical and Quantitative Biology, and Department of Biochemistry and Cell Biology, Renaissance School of Medicine, Stony Brook University, Stony Brook, NY, USA*
5. *University of Pittsburgh Hillman Cancer Center, Pittsburgh, PA, USA*
6. *Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, PA, USA*

[a] *Equal contribution*

[b] *Corresponding authors:* xing1@pitt.edu, bahar@laufercenter.org, mchen@zju.edu.cn.

Abstract

RNA velocities and generalizations emerge as powerful approaches for extracting time-resolved information from high-throughput snapshot single-cell data. Yet, several inherent limitations restrict applying the approaches to genes not suitable for RNA velocity inference due to complex transcriptional dynamics, low expression, or lacking splicing dynamics, or data of non-transcriptomic modality. Here, we present GraphVelo, a graph-based machine learning procedure that uses as input the RNA velocities inferred from existing methods and infers velocity vectors lying in the tangent space of the low-dimensional manifold formed by the single cell data. GraphVelo preserves vector magnitude and direction information during transformations across different data representations. Tests on multiple synthetic and experimental scRNA-seq data including viral-host interactome and multi-omics datasets demonstrate that GraphVelo, together with downstream generalized dynamo analyses, extends RNA velocities to multi-modal data and reveals quantitative nonlinear regulation relations between genes, virus and host cells, and different layers of gene regulation.

Introduction

Cells need to constantly detect and adapt to changes in extracellular and intracellular environments. Regulation of their gene transcription is a common mechanism of response. Multiple factors affect the transcriptional activity of eukaryotic genes, including *cis* and *trans* regulatory elements and chromatin structure. High throughput single-cell sequencing data provide the landscape of cell genotype. These data lack, however, information on how the cell state changes over time. Continuous efforts have been made to extract information about gene regulation and develop methods for connecting the cell states to temporal sequences of events captured by single-cell snapshot data. One group of methods that has received extensive attention is based on RNA velocity[1] for predicting the changes in RNA expression states in the cell. The original RNA velocity method leverages the ratio between nascent and mature transcripts to estimate the rate of change in gene expression. This seminal study has inspired numerous methods for improved RNA velocity estimation based on information from splicing[2-6], metabolic labeling[7,8], lineage tracing[9], and transcriptional factor binding[10], etc.

The RNA velocity framework has, however, its inherent limitations. First, none of the RNA velocity estimation methods could be applied to any single cell transcriptomic data without restrictions. For example, the splicing-based method is not applicable to prokaryotes or viruses, or organisms without introns. Erroneous inferences of RNA velocities have also been noticed for genes having complex splicing dynamics[11]. Furthermore, it is difficult to estimate the RNA velocities of genes with low expression, which excludes most transcription factors. Second, multi-omics sequencing technologies provide multifaceted information on cellular states alongside the transcriptome modality, and there exist limited systematic methods to extend such velocity estimations to other modalities[12,13].

The single cell transcriptome state is usually defined by the instantaneous distribution of RNA levels, represented by a multidimensional vector of RNA levels for all (measurable) genes. The usual practice is to map such single cell data onto a reduced space, e.g., transform from a principal component (PC) space to a UMAP representation to facilitate the visualization of the time-evolution of cell state, with each cell state being represented by a point in that space. Numerous dimension reduction and manifold learning algorithms have been developed for representation transformation. In comparison, transforming a velocity vector between representations is a

nontrivial task not rigorously addressed in the single cell field. Even worse, a visually correct vector field does not necessarily imply accurate high-dimensional velocity estimation[14]. La Manno et al. proposed a cosine kernel method to address this challenge, which has been adopted since then in most subsequent studies[1]. Li et al. mathematically proved that the cosine kernel asymptotically gives the correct direction of a velocity vector in the large sampling limit, but the magnitude information is completely lost due to a normalization procedure[15]. This loss of information casts concerns when such quantitative information is needed.

In this study, we tackle the above challenges through a graph-theoretical representation of RNA velocities, called GraphVelo, with dynamical systems underpinnings. GraphVelo takes an ansatz that the measured single cell expression profiles and inferred RNA velocities collectively reflect a dynamical process and are connected through a set of dynamical equations. It exploits such additional constraints that couple a high dimensional velocity field and a corresponding single cell state manifold, and enables the generalization of the approach in the context of multi-modal single cell data. While the combined expression and velocity information has been widely used to infer cell state transition trajectories[2,16], GraphVelo presents the advantage of enabling downstream analyses such as that performed by dynamo[7] to extract quantitative information on causal gene-gene relations that dictate the cell state transitions. Benchmarking of the proposed graph framework against simulated and experimental single cell data lends support to its broad utility.

Results

GraphVelo infers manifold-consistent single cell velocity vectors through tangent space projection and transforms between representations through local linear embedding.

Consider that the internal state of a cell can be specified by an N-dimensional state vector $\boldsymbol{x}$, with $N >> 1$ generally. Assume that the temporal evolution of the cell state follows a continuous and smooth curve $\boldsymbol{x}(t)$ (see Methods for more general mathematical formulation). The instant velocity vector $\boldsymbol{v}(\boldsymbol{x}, t) = \mathrm{d}\boldsymbol{x}(t)/\mathrm{d}t$ is always tangent to the curve of $\boldsymbol{x}(t)$ (as a function of t) at $\boldsymbol{x}$. One can generalize to the situation that the trajectories of a swarm of cells form a M-dimensional manifold $\mathcal{M}(\boldsymbol{x})$ embedded in the N-dimensional state space with $M << N$ typically, as revealed by high-throughput single cell omics data. Then under the ansatz that a velocity vector $\boldsymbol{v}(\boldsymbol{x}, t)$ dictates the evolution of a state vector $\boldsymbol{x}(t)$, $\boldsymbol{v}$ must lie in the tangent space of $\mathcal{M}(\boldsymbol{x})$, denoted as $T_p\,\mathcal{M}$. In

practice, the RNA velocity vectors inferred from existing methods do not automatically satisfy this tangent space requirement (but see[4]).

Taking various inferred single cell RNA velocity vectors, e.g. splicing-based, metabolic labeling-based, or lineage tracing-based, as input, GraphVelo takes advantage of the nature of the low-dimensional cell state manifold to: 1) refine the estimated RNA velocity to satisfy the tangent space requirement; 2) infer the velocities of non-transcriptomic modalities using RNA velocities. GraphVelo thus serves as a plugin that can be seamlessly integrated into existing RNA velocity analysis pipelines, and help process single cell data for downstream cellular dynamics analyses using methods such as dynamo (Fig. 1a).

Practically, GraphVelo approximates the tangent space at a cell state $\boldsymbol{x}$ by a k-nearest neighbor (kNN) graph following the local linear embedding algorithm[17], and uses the more reliable data manifold $\mathcal{M}$ to refine the velocity vectors by imposing the constraint that the N-dimensional velocity vector $\boldsymbol{v}$ should lie in the tangent space (Fig. 1b&c). Consider a given point $\boldsymbol{x}_i$ on a manifold corresponding to the expression state i of the single cell. Its infinitesimal neighborhood forms an Euclidean space that approximates the tangent space $T_p\mathcal{M}$. With a sufficient sampling of the neighboring cell states j in the state space, the incremental displacement vectors between cell state i and its neighboring cell states, $\boldsymbol{\delta}_{ij} = \boldsymbol{x}_j - \boldsymbol{x}_i$, form a set of complete albeit possibly redundant and nonorthogonal/non-normalized basis vectors of the Euclidean space in the local region. Then the projection of the measured velocity vector onto $T_p\mathcal{M}$ can be expressed as a linear combination (see also Supplemental Text I),

$$\boldsymbol{v}_{\parallel}(\boldsymbol{x}_i) = \sum_{j\in\mathcal{N}_i} \phi_{ij}\boldsymbol{\delta}_{ij}, \tag{1}$$

where $\mathcal{N}_i$ is the neighborhood of cell state i, defined by its k nearest neighbors in the feature space determined by sequencing profiles. Direct application of eq. 1 to determine the coefficients ϕ_{ij} is numerically unstable in real data (see Supplemental Text II for detailed discussion). Instead, we performed the projection by optimizing the following tangent space projection (TSP) loss function (Fig. 1c),

$$\mathcal{L}(\boldsymbol{\phi}_i) = a \cdot \left\|\boldsymbol{v}_i - \boldsymbol{v}_{\parallel i}\right\|^2 - b \cdot \cos(\boldsymbol{\phi}_i, \boldsymbol{\phi}_i^{corr}) + \lambda\|\boldsymbol{\phi}_i\|^2, \tag{2}$$

where $\|\cdot\|$ refers to vector modulus. $\boldsymbol{\phi}_i^{corr}$ is a heuristic "cosine kernel" widely used in the RNA velocity analyses for projecting velocity vectors onto a reduced space, the elements of which are

ϕ_{ij} $(with\ j \in \mathcal{N}_i)$(see also Supplemental Text III); the second term $\cos(\cdot,\cdot)$ denotes the cosine similarity. The first term in the loss function learns the correct velocity magnitudes, and the second term retains the reliable direction information based on previous mathematical analyses showing that $\boldsymbol{\phi}_i^{corr}$ asymptotically gives correct direction of the velocity vector[15]. The L2 regularization is used to bound parameters ϕ_i . Hyperparameters a, b and λ are for retaining the projection strength, direction, and for regularization, respectively.

With local linear embedding, it is straightforward to transform velocity between different representations. Assuming a mapping function f exists connecting manifold $\mathcal{M}$ and $\aleph$ such that for cell i with state vector $\boldsymbol{x}_i$ in $\mathcal{M}$, the coordinate of the same cell in $\aleph$ is given as $\boldsymbol{y}_i = f(\boldsymbol{x}_i)$. Since a given local patch of continuous manifold is approximated by an Euclidean space, a locally linear transformation connects the patch in the two representations. Consequently, for a vector described by eq. 1 in $\mathcal{M}$, the velocity vector in $\aleph$ is,

$$\boldsymbol{v}_{\parallel}(\boldsymbol{y}_i) = \sum_{j \in \mathcal{N}_i} \phi_{ij} \boldsymbol{\delta}_{ij}^{'}, \tag{3}$$

where $\boldsymbol{\delta}_{ij}' = \boldsymbol{y}_j - \boldsymbol{y}_i$. That is, one only needs to change the basis vectors.

Therefore, eq. 1-3 form the mathematical and computational foundation of GraphVelo. With eq. 3 one can extend velocity inference to datasets that velocity inference is not traditionally applicable such as host-virus interactome and multi-omics datasets based on the Whitney embedding theorem[18]. Details of the mathematical foundation were given in Methods. With velocity vectors refined with GraphVelo, one can readily perform downstream analyses, as exemplified in Fig.1d-f, which will be further elaborated below in the context of specific applications.

Benchmark studies demonstrate effectiveness of GraphVelo across simulation datasets with diverse topology

To demonstrate the effectiveness of the geometry-constrained projection, we first benchmarked our method on a 3D bifurcation system constrained on a 2D manifold (Methods). We added random components vertical to the tangent plane to mimic the noise. The resulting velocity vectors inferred by GraphVelo through minimizing the TSP loss were consistent with the ground truth vectors (Fig. 2a). Both GraphVelo and cosine kernel successfully removed the normal components (Fig. 2b i) and maintained the directional information (Fig. 2b ii), but only GraphVelo kept the velocity magnitude information (Fig. 2b iii).

155 Next, we performed multifaceted evaluations of the ability of GraphVelo to robustly recover the
156 transcriptional dynamics across a range of simulated datasets with different underlining phenotypic
157 structures. We used dyngen[19], a multi-modal scRNA-seq simulation engine, to generate gene-wise
158 dynamics defined by gold-standard transcriptional regulatory networks (Methods). We generated
159 simulated scRNA-seq data for networks with a variety of underlying linear, cyclic, and bifurcating
160 topological structures, and recovered the corresponding vector field using GraphVelo-corrected
161 velocity vectors (Fig. 2c-e). To comprehensively assess the outcome, we used three diverse
162 metrics, cosine similarity, root-mean-square error (RMSE), and accuracy, which evaluate the
163 correctness of velocity direction, magnitude, and sign, respectively. We presented (Fig. 2c-e) the
164 comparative results obtained with the cosine kernel and with GraphVelo (TSP with (i.e., eq. 2 with
165 $b \neq 0$) and without (eq. 2 with $b = 0$) the cosine regularization term). By minimizing TSP loss,
166 GraphVelo preserved both the direction and magnitude of the vector field (Fig. 2f-h). With an
167 increase of noise level by adding Gaussian noise to the ground truth vectors, GraphVelo refined
168 the distorted velocity and outperformed the cosine kernel projection consistently (Extended Data
169 Fig. 1a-c).

170 Then, we tested whether manifold constraints could preserve the speed of the cell progression
171 across different representations. GraphVelo was able to scale velocity vectors between the original
172 space and the PCA space, showing a high correlation with the ground truth, even as noise levels
173 increased, whereas the cosine kernel failed (Extended Data Fig. 1d). The results on UMAP showed
174 less agreement, which is not surprising. UMAP is a convenient representation for visualizing single
175 cell data but is not designed for representing quantitative cell state transition dynamics. That is,
176 UMAP is not a continuous transformation from the original gene space and cannot preserve local
177 distances after projection.

178 To further explore whether GraphVelo could correct the RNA velocity estimated by the splicing
179 kinetics, we took the velocity inferred using different packages (scVelo[2], dynamo[7] and VeloVI[5])
180 as input. The output from GraphVelo agreed significantly better with the ground truth compared
181 to the raw input (Extended Data Fig. 1e), highlighting the significant improvement achieved by
182 GraphVelo in evaluating both the direction and magnitude of the velocity vector fields across all
183 datasets.

184 **GraphVelo achieved consistent improvements across multiple real-world datasets**

185 Although GraphVelo is designed as a velocity-correction and prediction model complementing
186 existing velocity estimation tools, we benchmarked the performance of scVelo-based GraphVelo
187 outputs on five independent datasets (FUCCI[20], pancreatic endocrinogenesis[21], dentate gyrus[1],
188 intestinal organoid[22] and hematopoiesis[7]) against five RNA velocity estimation methods (scVelo[2],
189 VeloVI[5], UniTVelo[23], DeepVelo[24] and CellDancer[3]) (Extended Data Fig. 2a). GraphVelo achieved
190 noticeably improved cross-boundary correctness (CBC) score[25] against input velocity and other
191 advanced methods (see Methods for calculation details).

192 We examined the cell cycle datasets annotated by the dynamo package, which features a relatively
193 simple geometry, to quantitatively evaluate our method. First, we focused on the CBC score
194 between cell cycle states. The velocity vectors processed by TSP showed greater consistency with
195 the ground truth compared to the unprocessed inputs (Extended Data Fig. 2b). To demonstrate the
196 effectiveness of GraphVelo in scaling velocities based on the data manifold, we used the L2 norm
197 of velocity vectors to quantify the cell cycle speed. This analysis revealed a peak in velocity within
198 the M and G1 phases, which was also reflected in the distribution of total UMI counts (Extended
199 Data Fig. 2c). We further validated these findings using the cycling A549 cell line sequenced by
200 sci-fate[26] (Extended Data Fig. 2d) and through the temporal variation of stratified cell cycle speed
201 based on velocities inferred from metabolic-labeling data (Extended Data Fig. 2e). Leveraging the
202 quantitative velocity vectors generated by GraphVelo, we classified genes by both the phase and
203 peak magnitude of their velocities. Analysis of phase-magnitude relationships uncovered the
204 sequential activation cascade of marker genes throughout the cell cycle.

205 **GraphVelo infers quantitative genome-wide RNA velocity from a subset of genes with**
206 **manifold-consistent RNA turnover kinetics**

207 Most RNA velocity methods are based on biophysical models of mRNA turnover dynamics with
208 specific assumptions that may break down in certain cases[11]. These methods typically provide
209 velocities of only a subset of (~ 500 or less) genes, termed velocity genes in the subsequent
210 discussions, out of a larger list of (~2-3 k) highly variable genes in a dataset, and some of the
211 velocities are questionable. For example, the splicing-based RNA velocity may have an erroneous
212 sign for processes under active regulation on mRNA degradation or promotors switching between
213 states with different transcription efficiency (Fig. 3a, Extended Data Fig. 3).

214 GraphVelo uses the velocities of high-confidence genes obtained from any method as input to infer
215 velocities of other genes. One can use several existing approaches to evaluate the confidence scores
216 of inferred RNA velocity values of genes[7]. Alternatively, we identified a subset of Manifold-
217 consistent Kinetics (MacK) Genes based on their agreement with prior knowledge or additional
218 information acquired from other methods such as lineage tracing (Fig. 3b), We first applied
219 GraphVelo to a mouse erythroid maturation dataset[27]. This study provided a transcriptional
220 landscape of the erythroid lineage with well-documented differentiation trajectory during mouse
221 gastrulation. Previous analyses have shown that the dataset contains genes with multiple rate
222 kinetics, leading to erroneous prediction of the cell state transition direction[27,28]. We selected the
223 top 200 out of 450 velocity genes as MacK genes, representing those with robustly estimated
224 velocities (see Methods for details). The projected vector field in UMAP showed consistency with
225 prior knowledge in developmental biology (Fig. 3c). We then used the corrected RNA velocities
226 for dynamo velocity field analyses. The vector field-based pseudotime accurately predicted the
227 lineage with scRNA-seq data of temporal mouse embryos (Fig. 3d).

228 Previous studies identified multiple rate kinetics (MURK) genes showing transcription bursts in
229 the middle of erythroid differentiation[28]. For example, two MURK genes, *Smim1* and *Hba-x*,
230 showed complex patterns of phase portrait (Fig. 3f). Consequently, the RNA velocity of *Simi1*
231 inferred with scVelo was negative along a major part of the developmental axis (Fig. 3g i),
232 contradicting the trend of increasing *Simi1* mRNA levels (Fig. 3g iii). For *Hba-x*, scVelo even
233 failed to infer its RNA velocity. On the other hand, GraphVelo inferred velocities and predicted
234 correct kinetic patterns of these genes (Fig. 3g ii). Similar performances have been observed in
235 other MURK genes (Extended Data Fig. 4a). To examine the overall prediction of cell state
236 transitions from transcription burst genes, we projected he MURK genes velocity inferred from
237 GraphVelo and scVelo to the predefined UMAP. The velocities from GraphVelo but not scVelo
238 correctly captured the directional flow of differentiation using only MURK genes (Extended Data
239 Fig. 4b). We further recapitulated gene succession and oscillation magnitudes along the erythroid
240 trajectory and evaluated the phase-magnitude relationships of all highly variable genes (Extended
241 Data Fig. 5a). Compared to non-MURK genes, MURK genes exhibited larger average velocity
242 magnitudes and were predominantly enriched in the late stages of lineage progression. Analysis of
243 genes with larger peak velocity amplitudes identified *Fth1*, *Car2*, and *Hbb-bs* as candidates with
244 altered kinetic parameters. These dynamics patterns were evident in their phase portraits and gene

245 expression trends (Extended Data Fig. 5b&c), consistent with previous reports that *Car2*
246 transcription in erythroid cells is regulated by both the promoter activity and long-range enhancer
247 interactions[29]—such complex regulations lead to its transcription dynamics not well-described by
248 the simple transcription model used in the original splicing -based RNA velocity inference.

249 Next, we evaluated the quantitative performance of GraphVelo in estimating cell-wise transition
250 speed. Specifically, we estimated the speed of cell state transition using the norm of velocity vector
251 in high-dimensional space and identified the transcriptional surge stage (Extended Data Fig. 6c).
252 We hypothesized that the MURK genes, which exhibited a sudden increase in transcription rate
253 during this stage, were responsible for the sharp acceleration in cell state transition speed. Using
254 dynamo, we estimated the acceleration derived from the GraphVelo vector field and found that the
255 acceleration value, as the derivative of the velocity vector, demonstrated its potential as a predictor
256 for transcription burst genes (Extended Data Fig. 6d).

257 With the velocity estimation extended to the whole gene space, we were able to perform
258 comprehensive mechanistic analyses on the entire genome spectrum. First, we calculated the
259 MacK score for each gene using the corrected RNA velocities. We hypothesized that a gene with
260 a higher MacK score indicated a better agreement between its RNA velocity vector and the
261 developmental axis, suggesting that the gene served as a potential lineage-driver gene. We ranked
262 genes based on their scores and performed GO biological process enrichment analyses for the top
263 genes. Indeed, the enriched processes were associated with erythropoiesis, including the heme
264 biosynthetic process and interleukin-12-mediated signaling pathway[30] (Fig. 3e).

265 Next, we applied dynamo to perform differential geometry analyses of the vector field and
266 mechanistically dissected the activation cascade of erythroid marker gene *Klf1* (Extended Data Fig.
267 6a&b). Jacobian analyses based on GraphVelo vector field revealed sequential activation of driver
268 transcription factors (TFs) *Gata2*, *Gata1*, and *Klf1* during erythroid lineage differentiation, with
269 *Gata1* subsequently repressing the expression of *Gata2* (Fig. 3h, Extended Data Fig. 6c)[7]. To
270 further demonstrate the crucial role of transcriptional factor *Gata1* during erythropoiesis, we
271 performed *in silico* genetic perturbation across all cells. Results showed that both inhibiting *Gata1*
272 and upregulating the *Gata1* repressor *Spi1* lead to a reversal of normal developmental flow (Fig.
273 3i, Extended Data Fig. 6d). The above analyses collectively suggest that activation of *Gata1* in the
274 blood progenitors biased its differentiation to erythropoiesis, agreeing with experimental reports[28].

275 To further evaluate GraphVelo, we tested the method on another dataset of human bone marrow
276 development[31]. This developmental process has complex progressions from hematopoietic stem
277 cells (HSCs) to three distinct branches: erythroid, monocyte, and common lymphoid progenitor
278 (CLP). Again, we used the top 100 out of 454 velocity genes as MacK genes to predict the RNA
279 velocities of 2,000 highly variable genes. The GraphVelo velocity field accurately recovered the
280 fate of cells on the sophisticated transcriptional landscape in contrast to scVelo (Fig. 3j and k,
281 Extended Data Fig. 7a). By combining the likelihood estimated by scVelo with the MacK score,
282 we identified rapid degradation and transcription burst genes whose dynamics deviated from the
283 RNA velocity assumptions (Extended Data Fig. 7b&c). *ANGPT1* and *RBPMS* are two examples
284 which were overall highly expressed in the progenitors and decreased quickly along the trajectories
285 (Fig. 3l), reminiscent of what was shown in Fig. 3a. These genes misled RNA velocity inference
286 with scVelo assuming a constant degradation rate constant. GraphVelo revealed a cell context-
287 specific transcription rate $\alpha = u + \frac{du}{dt}$ and degradation constant $\gamma = (u - \frac{ds}{dt})/s$, thus a
288 degradation wave along the differentiation path (Fig. 3l, Extended Data Fig. 7d), consistent with
289 simulation result and reports on regulation of *ANGPT1* mRNA by microRNAs such as *miRNA-*
290 *153-3p*[32] (Extended Data Fig. 2d).

291 **GraphVelo reconstructs host–pathogen transcriptome dynamics from infection trajectory**

292 Continuous battle between human immune surveillance and viral immune evasion takes place in
293 the host cell system after viral entry. scRNA-seq data provide a massive and parallel way of
294 assessing the time evolution of both host and viral transcripts, unraveling the delicate inherent
295 dynamics of a virus-host system[33,34]. For splicing-based RNA velocity methods, genes lacking
296 sufficient unspliced transcript counts are typically filtered out—which is inherently the case for
297 viral genes due to their absence of introns. However, several studies[35-37] have applied tools such
298 as scVelo to host-virus systems by focusing exclusively on host velocity genes that accurately
299 reflect the lineage dynamics, rather than attempting to derive velocities directly from viral
300 transcripts. GraphVelo enables inference of the velocities of virus RNA abundance based on the
301 kinetics of host transcripts velocities, as illustrated next.

302 We analyzed a human cytomegalovirus (HCMV) viral infection dataset to learn viral
303 transcriptomic kinetics in monocyte-derived dendritic cells (moDCs)[37]. The result from
304 GraphVelo unraveled how viral infection progressed along the transcriptional space (Fig. 4a). The

305 velocity vectors pointed to directions consistent with an increasing trend of the percentage of viral
306 RNAs in individual cells, which inherently served as an indicator of the infection time course[38].
307 Compared to the trend obtained with the raw RNA velocities from scVelo, the vector field-based
308 pseudotime calculated using GraphVelo-corrected RNA velocities consistently showed higher
309 correlation with the (pseudo)temporal progression of viral infection as reflected by viral RNA
310 percentage (Fig. 4b).

311 Furthermore, the examination of individual genes revealed that the RNA velocities consistently
312 predicted the trend of the mRNA expression level change with increasing virus load (Fig. 4c,
313 Extended Data Fig. 8a&b). Most viral genes started with a fast-increase phase, and the expressions
314 of some genes (e.g. *UL22A*) gradually saturated at high virus load, together with the corresponding
315 RNA velocities approaching zero. One exception is *UL122*, whose expression profile increased
316 first then decreased to a steady state level lower than the peak value. This overshooting is
317 characteristic of a negative feedback network structure[39]. Indeed, a recent study reported that
318 *UL122* negatively regulates its own promotor[40]. Furthermore, comparison of MacK scores across
319 GraphVelo, the CellRank pseudotime kernel, and randomized prediction showed that GraphVelo-
320 computed viral RNA velocities aligned best with the transcriptome gradient of viral load (Fig. 4d).
321 Note that the MacK score also served as a reliable predictor for dynamics-driving factors,
322 specifically viral genes in this case (Extended Data Fig. 8d).

323 We further quantified the velocity norm of all viral factors as infection speed, and observed that
324 the transcription of viral factors was significantly restricted initially, then gradually increased
325 along the trajectory (Extended Data Fig. 8e). Interestingly, most of the genes that exhibited positive
326 correlation with the infection speed were related to viral DNA synthesis, while those negatively
327 correlated to the infection speed were engaged in host viral defense response[41](Fig. 4e, Extended
328 Data Fig. 8f).

329 **GraphVelo identifies host genomic response modules and predicts host-virus gene**
330 **interactions**

331 With GraphVelo-inferred RNA velocities, we probed the time evolution of lytic infection and the
332 complex interplay between host and viral functional genomes. By fitting the GraphVelo velocity
333 trends along the viral load axis, we identified genes with similar kinetic patterns (Fig. 4c). Using
334 the smoothened velocity trends to calculate the distances, we clustered the genes into seven major

335 modules and visualized them on the UMAP space (Fig. 4f). Not surprisingly, viral genes were
336 concentrated in several enclosed regions, indicating that they formed distinct functional genomic
337 modules during the lytic cycle[38]. The genes showed two major dynamical features: acceleration
338 and deceleration along the viral load axis (Fig. 4g).

339 To systematically investigate whether the dichotomy between host kinetics genes and viral genes
340 share similar dynamics, we performed gene functional enrichment analyses of host genes residing
341 close to the viral gene clusters. Genes located in the deceleration part were associated with
342 repressing viral genome replication, which includes negative regulation of viral life cycle and
343 known restriction factors in antiviral responses activated in DCs such as the induction of cytokine
344 and chemokine responses as well as interactions with neutrophils[42,43]. In parallel, the toll-like
345 receptor signaling pathway, required for antiviral defense of the host, was arrested[44]. Neutrophil
346 related processes, which typically cooperate closely with DCs to modulate adaptive immune
347 responses[45], were suppressed. Therefore, the deceleration part also showed how critical set of host
348 factors were silenced by viral entry to achieve immune evasion.

349 The acceleration groups, on the other hand, demonstrated how viruses hijacked the host cell
350 endogenous cellular programs for virus replication. Notably, the pathways related to viral genome
351 replication were triggered, promoting DNA replication and transcription, such as negative
352 regulation of G1/S transition of mitotic cell cycle[46], cellular response to DNA damage stimulus[47]
353 and regulation of transcription from RNA polymerase II promoter[48]. Cells showed a shift towards
354 a transcriptional signature resembling the G1 phase (Extended Data Fig. 8g), agreeing with
355 previous report on HCMV infection[49]. Along the infection process, antiviral interferon (IFN)-γ
356 response of moDC cells was first activated then suppressed. These results highlighted the
357 organized and antagonistic strategies adopted by both host cells and viruses during their tug-of-
358 war for survival and proliferation.

359 To investigate the crosstalk between host and viral factors systematically in depth, we performed
360 dynamo Jacobian analyses. We scanned the entire spectrum of viral genome and delineated how
361 the HCMV factors silence IFN and NFκB signaling (Fig. 4i). A large proportion of the identified
362 viral factors functioned in evading host cell immune responses, a finding supported by several
363 recent studies[50-52]. *In silico* virus-directed knock out experiments revealed altered accumulation
364 patterns of viral transcripts (Fig. 4j). Notably, inhibition of *UL123*, which ranked first with the

365 total viral RNA inhibition in our analyses, led to a qualitatively distinct trajectory. These results
366 highlight the multifunctional *UL123* locus in the viral genome as a potential target for antiviral
367 intervention[38,53]. The analyses demonstrated potential usage of GraphVelo-inferred velocities for
368 understanding the interactions between viral and host factors, assessing the effects of perturbations
369 on infection, and designing potential antiviral interventions[38].

370 **GraphVelo reaveals an abortively infected cell population in SARS-CoV-2 infection**

371 Although HCMV is renowned for its elaborate transcriptional landscape—characterized by
372 extensive alternative splicing, overlapping transcripts, and diverse isoform expression—the RNA
373 virus SARS-CoV-2 transcriptome reveals an even greater level of complexity within its relatively
374 compact ~30 kb genome[54]. To characterize the molecular mechanisms of host response which
375 protect cells from productive trajectory, we applied GraphVelo to a SARS-CoV-2 infected Calu-
376 3 cells dataset[55]. To focus on the host-virus interactions and their corresponding fate outcomes, we
377 subset the infected cell population for downstream analyses. The infected cluster M, characterized
378 by interferon production genes, was hypothesized to represent a subpopulation of abortively
379 infected cells[55], similar to those described in herpesviruses HSV-1[56] and HCMV[38]. We subsampled
380 the infected cell population for downstream analyses and visualization, where cluster M exhibited
381 distinct connectivity properties compared to the main infected groups (Extended Data Fig. 9a). To
382 confirm that M is one of the terminal states of infection outcomes with such low cell abundance
383 (Extended Data Fig. 9b), we applied GraphVelo to gain the quantitative velocity with dynamo
384 outcomes as input (Extended Data Fig. 9c). Using vector field topology analysis, we classified an
385 initial state with relatively low viral load, two terminal states associated with high apoptosis
386 activity, and a saddle point characterized by high viral load. Notably, the region corresponding to
387 cluster M was identified as an attractor, confirming it represents an abortively infected cell state
388 with a high death rate[55].

389 We further validated the complex lineage commitments of SARS-CoV-2–infected cells using the
390 CellRank framework[16], which successfully identified three potential terminal states as outcomes
391 of host-virus competition (Extended Data Fig. 9e). Interestingly, we characterized the saddle stage
392 in dynamo as a productive terminal state governed by viral genes (Extended Data Fig. 9f),
393 exhibiting high viral transcription speed (Extended Data Fig. 9g). The pathogen-triggered cell
394 death can be driven for protecting the host or for pathogen dissemination purpose[57]. To investigate

395 the underlying causes for host cell death, we performed gene functional enrichment analyses on
396 the top correlated genes along distinct lineages (Extended Data Fig. 9h). Active programs in the
397 abortive infection lineage were highly enriched for host defense mechanisms against viral invasion,
398 consistent with previous studies[38,56]. In contrast, the drivers of the apoptosis-associated lineage
399 revealed a distinct profile. Notably, we observed enrichment for cellular responses to unfolded
400 proteins, which have been implicated in facilitating pathogen-mediated dissemination of infected
401 cells[58,59]. Additionally, *ERBB2* inhibition has been shown to suppress SARS-CoV-2 replication[60].
402 These findings support the hypothesis that similar cell death outcomes may arise from
403 fundamentally different host responses. Abortively infected cells appear to promote efficient
404 pathogen clearance, likely through cytokine-mediated immune activation that eliminates both the
405 infected cell and the virus. Conversely, cells undergoing virus-induced apoptosis fail to clear the
406 virus, with cell death instead serving as a mechanism for viral escape, immune evasion, and
407 potential dissemination to deeper tissue layers or the bloodstream.

408 **GraphVelo permits multi-omics velocity inference and chromatin dynamics analyses**

409 The molecular anatomy during cell development entails multiple layers, and how different layers
410 coordinate to regulate gene expression is a fundamental problem. For example, the anagen hair
411 follicle features distinct lineages branching from a central population of progenitor cells. Ma et
412 al[61]. used SHARE-seq to capture both the transcriptome and the epigenome data simultaneously
413 for the lineage commitment process from transit-amplifying cells (TACs) to the inner root sheath
414 (IRS), cuticle layer, and medulla. Upon robust selection of estimated genes following dynamo
415 criteria (Methods), we further refined the RNA velocities of these genes through tangent space
416 projection and obtained the chromatin open/close dynamics from the corresponding scATAC data
417 using GraphVelo. The resultant vector field in the combined transcriptome-epigenome space
418 proved to reconstruct the correct multilineages differentiation paths during the anagen phase (Fig.
419 5a).

420 To test the consistency of dynamics across different modalities, we performed CellRank terminate
421 stage analyses[16] from the refined velocity vectors. Using GraphVelo velocities of either the RNA
422 modality or the ATAC modality, we accurately estimated three diverse terminal stages (Fig. 5b).
423 For comparison, we also performed similar analyses using MultiVelo, scVelo with all velocity
424 genes or robustly estimated genes in above GraphVelo studies and pseudotime-based vector field

425 inferred by CellRank. The 2D projection of these vector field functions also exhibited seemingly
426 correct velocity flow direction (Extended Data Fig. 10a). However, none of them captured the cell
427 fate commitment based on coarse-grained transition matrix (Fig. 5b, Extended Data Fig. 10b).
428 Notably, the results from the RNA modality and the ATAC modality of MultiVelo gave
429 inconsistent results. GraphVelo-corrected velocities, on the other hand, helped identify the top-
430 correlating genes towards individual terminal populations which showed agreement with previous
431 study[62] (Fig. 5c, Extended Data Fig. 11).

432 Next, we conducted differential geometry analyses based on the composite GraphVelo vector field.
433 We identified novel root cells, which were also characterized by chromatin potential (Fig. 5d)[61].
434 These novel root cells expressed distinct marker genes compared to the expected root cells using
435 the Wilcoxon test (Fig. 5e). Moreover, we unraveled differentially expressed markers identified
436 by the original study[61], as well as new differentiation-potent genes and validated their initiation
437 properties in another transcriptome dataset (Extended Data Fig. 12a)[62]. To further investigate how
438 these two distinct groups of root cells convert to other cell types, we performed a least action path
439 (LAP) analysis between different cell phenotypes. The expected and novel root cells converted to
440 the IRS terminal state following two distinct least action paths in the vector field (Extended Data
441 Fig. 12b&c). The two paths revealed different temporal change patterns of transcription factor
442 expression profiles (Extended Data Fig. 12d). We calculated the mean-squared displacement
443 (MSD) for every transcription factor to explore the dynamics of TFs along the path from novel
444 root to IRS. The result demonstrated that the fate conversion by novel root was mediated by the
445 *Shh-Runx1* signaling axis (Fig. 5f, Extended Data Fig. 12d), which has been demonstrated in
446 human embryonic stem cells[63] and is crucial for hair development[64]. In summary, GraphVelo
447 unraveled the multiple molecular mechanisms that orchestrated hair follicle morphogenesis.

448 With available chromatin velocity and RNA velocity, we set to quantify the coupling/decoupling
449 relationships between chromatin structure and gene expression for each gene (see Methods). Here
450 chromatin structure refers to the extent of exposure/accessibility of the gene locus to the
451 environment as indicated by scATACseq data, shortly called open or closed state; and chromatin
452 velocities refer to changes between these states as inferred from scATACseq counts at the
453 examined gene locus. We used dynamic time warping (DTW) distances between the velocities
454 from different omics layers to quantify the similarity between temporal patterns of these two

modalities for each gene. A higher DTW value indicates higher similarity. Using the elbow of the ranked distance curve as a cutoff we identified genes that showed decoupled transcription and chromatin structure dynamics. These decoupled genes had an accumulation of cell cycle-dependent (CCD) genes found in previous study[20] (Fig. 5g). This group of genes showed strong involvement in cell cycle-related processes, as indicated by GO enrichment analyses (Fig. 5h). Close examinations indicated that the transcription of cell cycle related genes decreased along the differentiation path, while the chromatin structure at the corresponding loci remained open (Fig. 5i, Extended Data Fig. 13). To validate this hypothesis, we further applied GraphVelo to a recently published 10x Multiome dataset from developing human cortex[65] (Extended Data Fig. 14a). Following the same analyses, we identified decoupled genes and found out that most of these genes were related to cell cycle (Extended Data Fig. 14b-f). which has also been reported in a previous MultiVelo study[13].

We further performed dynamo differential geometry analyses on the composite transcriptome-chromatin vector field. One intriguing phenomenon observed in lineage dynamics is that *Lef1* and *Hoxc13* are the driver TFs correlated with domains of regulatory chromatin (DORCs) of *Wnt3* [61]. Differential geometry analyses on the composite vector field can go beyond correlation analyses and provide an underlying casual mechanism. As a prerequisite for such analyses, GraphVelo-inferred RNA and chromatin velocities of the three genes correctly predicted the trend of change of mRNA and ATAC-seq counts (Extended Data Fig. 15a&b), in contrast to the performance of MultiVelo (Extended Data Fig. 15c). Then, Jacobian analyses on the GraphVelo vector field confirmed that priming activation of *Lef1* subsequently activated the *Hoxc13* TF[66] (Extended Data Fig. 15d&e). Both *Lef1* and *Hoxc13* were found to activate the *Wnt3* target gene, initiating lineage commitment (Extended Data Fig. 15f &g). To quantitatively understand how the two TFs affect *Wnt3* chromatin structure and transcription, we plotted the response heatmap to reflect the distributions of Jacobian elements versus the abundance of mature mRNA for each TF (Fig. 5j). The two terms $\partial f_{Wnt3-\mathrm{chrom}} / \partial x_{Lef1}$ and $\partial f_{Wnt3-\mathrm{chrom}} / \partial x_{Hoxc13}$ started with positive values at low concentrations of TF mRNA copy numbers then decreased to zero, indicating that increasing the level of either TF lead to further opening of the *Wnt3* chromatin region, and the effect saturated at high TF expression. The other two term $\partial f_{Wnt3} / \partial x_{Lef1}$ and $\partial f_{Wnt3} / \partial x_{Hoxc13}$ increased with the TF levels, indicating that these two TFs also activated *Wnt3* transcription. Upon integration of the

Jacobian elements over regulator expression changes, we obtained the effective dose-response curves obtained (see Methods), which revealed more transparently the TF dose lag between the opening of the target chromatin region and the initialization of transcription (Fig. 5k).

Our results therefore illustrated the sequential events that these two driver TFs, *Lef1 and Hoxc13*, drove as pioneer transcription factors (PTFs) to initiate local chromatin opening and then activated the transcription of *Wnt3*. Notably, computational methods and experimental research confirm that *Lef1* acts as a nucleosome binder and exhibits diverse binding patterns across various cell lines[67]. The *Hox* family of TFs has also been shown to have the capacity to bind their targets in an inaccessible chromatin context and trigger the switch to an accessible state[68], consistent with our analyses that *Hoxc13* revealing that it shared a regulation mechanism similar to that of PTF *Lef1*.

Discussion

In this work, we provided a general framework that extends the framework developed for RNA velocity and related approaches to various data modalities such as proteomics, spatial genomics, 3d genome organization, and imaging data, which were originally beyond reach of this framework. We validated GraphVelo using various *in vivo* cellular kinetics models, confirmed its efficacy and robustness in handling complex and noisy multimodal data. Upon application to various datasets, we unraveled gene regulation relations of an extended list of genes, host-virus gene regulations, and coupling between transcription and local chromatin structures. GraphVelo can be seamlessly integrated with broad downstream analyses, such as dynamo continuous vector field analyses, as well as Markovian analyses using graph dynamo or CellRank.

References

1. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494-498 (2018).
2. Bergen, V., Lange, M., Peidli, S., Wolf, F.A. & Theis, F.J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol* **38**, 1408-1414 (2020).
3. Li, S. *et al.* A relay velocity model infers cell-dependent RNA velocity. *Nat Biotechnol* **42**, 99-108 (2024).
4. Lederer, A.R. *et al.* Statistical inference with a manifold-constrained RNA velocity model uncovers cell cycle speed modulations. *Nat Methods* (2024).
5. Gayoso, A. *et al.* Deep generative modeling of transcriptional dynamics for RNA velocity analysis in single cells. *Nat Methods* **21**, 50-59 (2024).
6. Gao, C.F., Vaikuntanathan, S. & Riesenfeld, S.J. Dissection and integration of bursty transcriptional dynamics for complex systems. *Proc Natl Acad Sci U S A* **121**, e2306901121 (2024).
7. Qiu, X. *et al.* Mapping transcriptomic vector fields of single cells. *Cell* **185**, 690-711 e45 (2022).
8. Weiler, P., Lange, M., Klein, M., Pe'er, D. & Theis, F. CellRank 2: unified fate mapping in multiview single-cell data. *Nat Methods* **21**, 1196-1205 (2024).
9. Wang, K. *et al.* PhyloVelo enhances transcriptomic velocity field mapping using monotonically expressed genes. *Nat Biotechnol* **42**, 778-789 (2024).

10. Li, J., Pan, X., Yuan, Y. & Shen, H.B. TFvelo: gene regulation inspired RNA velocity estimation. *Nat Commun* **15**, 1387 (2024).
11. Bergen, V., Soldatov, R.A., Kharchenko, P.V. & Theis, F.J. RNA velocity-current challenges and future perspectives. *Mol Syst Biol* **17**, e10282 (2021).
12. Gorin, G., Svensson, V. & Pachter, L. Protein velocity and acceleration from single-cell multiomics experiments. *Genome Biol* **21**, 39 (2020).
13. Li, C., Virgilio, M.C., Collins, K.L. & Welch, J.D. Multi-omic single-cell velocity models epigenome-transcriptome interactions and improves cell fate prediction. *Nat Biotechnol* **41**, 387-398 (2023).
14. Zheng, S.C., Stein-O'Brien, G., Boukas, L., Goff, L.A. & Hansen, K.D. Pumping the brakes on RNA velocity by understanding and interpreting RNA velocity estimates. *Genome Biol* **24**, 246 (2023).
15. Tiejun Li, T.L., Jifan Shi, J.S., Yichong Wu, Y.W. & Peijie Zhou, P.Z. On the Mathematics of RNA Velocity I: Theoretical Analysis. *CSIAM Transactions on Applied Mathematics* **2**, 1-55 (2021).
16. Lange, M. *et al.* CellRank for directed single-cell fate mapping. *Nat Methods* **19**, 159-170 (2022).
17. Roweis, S.T. & Saul, L.K. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* **290**, 2323-2326 (2000).
18. Whitney, H. The Self-Intersections of a Smooth n-Manifold in 2n-Space. *Annals of Mathematics* **45**, 220-246 (1944).
19. Cannoodt, R., Saelens, W., Deconinck, L. & Saeys, Y. Spearheading future omics analyses using dyngen, a multi-modal simulator of single cells. *Nat Commun* **12**, 3942 (2021).
20. Mahdessian, D. *et al.* Spatiotemporal dissection of the cell cycle with single-cell proteogenomics. *Nature* **590**, 649-654 (2021).
21. Bastidas-Ponce, A. *et al.* Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development* **146**(2019).
22. Battich, N. *et al.* Sequencing metabolically labeled transcripts in single cells reveals mRNA turnover strategies. *Science* **367**, 1151-1156 (2020).
23. Gao, M., Qiao, C. & Huang, Y. UniTVelo: temporally unified RNA velocity reinforces single-cell trajectory inference. *Nat Commun* **13**, 6586 (2022).
24. Cui, H. *et al.* DeepVelo: deep learning extends RNA velocity to multi-lineage systems with cell-specific kinetics. *Genome Biol* **25**, 27 (2024).
25. Qiao, C. & Huang, Y. Representation learning of RNA velocity reveals robust cell transitions. *Proc Natl Acad Sci U S A* **118**(2021).
26. Cao, J., Zhou, W., Steemers, F., Trapnell, C. & Shendure, J. Sci-fate characterizes the dynamics of gene expression in single cells. *Nat Biotechnol* **38**, 980-988 (2020).
27. Pijuan-Sala, B. *et al.* A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490-495 (2019).
28. Barile, M. *et al.* Coordinated changes in gene expression kinetics underlie both mouse and human erythroid maturation. *Genome Biol* **22**, 197 (2021).
29. Lee, J., Krivega, I., Dale, R.K. & Dean, A. The LDB1 Complex Co-opts CTCF for Erythroid Lineage-Specific Long-Range Enhancer Interactions. *Cell Rep* **19**, 2490-2502 (2017).
30. Dybedal, I., Larsen, S. & Jacobsen, S.E. IL-12 directly enhances in vitro murine erythropoiesis in combination with IL-4 and stem cell factor. *J Immunol* **154**, 4950-5 (1995).
31. Setty, M. *et al.* Characterization of cell fate probabilities in single-cell data with Palantir. *Nat Biotechnol* **37**, 451-460 (2019).
32. Ning, W. *et al.* Blocking exosomal miRNA-153-3p derived from bone marrow mesenchymal stem cells ameliorates hypoxia-induced myocardial and microvascular damage by targeting the ANGPT1-mediated VEGF/PI3k/Akt/eNOS pathway. *Cell Signal* **77**, 109812 (2021).
33. Bost, P. *et al.* Host-Viral Infection Maps Reveal Signatures of Severe COVID-19 Patients. *Cell* **181**, 1475-1488 e12 (2020).
34. Ratnasiri, K., Wilk, A.J., Lee, M.J., Khatri, P. & Blish, C.A. Single-cell RNA-seq methods to interrogate virus-host interactions. *Semin Immunopathol* **45**, 71-89 (2023).
35. Wilk, A.J. *et al.* A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nat Med* **26**, 1070-1076 (2020).
36. Edahiro, R. *et al.* Single-cell analyses and host genetics highlight the role of innate immune cells in COVID-19 severity. *Nat Genet* **55**, 753-767 (2023).
37. Costa, B. *et al.* Human cytomegalovirus exploits STING signaling and counteracts IFN/ISG induction to facilitate infection of dendritic cells. *Nat Commun* **15**, 1745 (2024).

38. Hein, M.Y. & Weissman, J.S. Functional single-cell genomics of human cytomegalovirus infection. *Nat Biotechnol* **40**, 391-401 (2022).
39. Alon, U. *An introduction to systems biology: design principles of biological circuits*, (Chapman & Hall/CRC, 2007).
40. Ball, C.B. *et al.* Human Cytomegalovirus IE2 Both Activates and Represses Initiation and Modulates Elongation in a Context-Dependent Manner. *mBio* **13**, e00337-22 (2022).
41. Manska, S. & Rossetto, C.C. Identification of cellular proteins associated with human cytomegalovirus (HCMV) DNA replication suggests novel cellular and viral interactions. *Virology* **566**, 26-41 (2022).
42. Gredmark-Russ, S. & Soderberg-Naucler, C. Dendritic cell biology in human cytomegalovirus infection and the clinical consequences for host immunity and pathology. *Virulence* **3**, 621-34 (2012).
43. Paulus, C., Krauss, S. & Nevels, M. A human cytomegalovirus antagonist of type I IFN-dependent signal transducer and activator of transcription signaling. *Proc Natl Acad Sci U S A* **103**, 3840-5 (2006).
44. Park, A. *et al.* HCMV-encoded US7 and US8 act as antagonists of innate immunity by distinctively targeting TLR-signaling pathways. *Nat Commun* **10**, 4670 (2019).
45. Yang, D., de la Rosa, G., Tewary, P. & Oppenheim, J.J. Alarmins link neutrophils and dendritic cells. *Trends Immunol* **30**, 531-7 (2009).
46. Bogdanow, B. *et al.* Human cytomegalovirus tegument protein pp150 acts as a cyclin A2-CDK-dependent sensor of the host cell cycle and differentiation state. *Proc Natl Acad Sci U S A* **110**, 17510-5 (2013).
47. Luftig, M.A. Viruses and the DNA Damage Response: Activation and Antagonism. *Annu Rev Virol* **1**, 605-25 (2014).
48. Ball, C.B. *et al.* Human Cytomegalovirus Infection Elicits Global Changes in Host Transcription by RNA Polymerases I, II, and III. *Viruses* **14**(2022).
49. Bogdanow, B., Phan, Q.V. & Wiebusch, L. Emerging Mechanisms of G(1)/S Cell Cycle Control by Human and Mouse Cytomegaloviruses. *mBio* **12**, e0293421 (2021).
50. Goodwin, C.M., Ciesla, J.H. & Munger, J. Who's Driving? Human Cytomegalovirus, Interferon, and NFkappaB Signaling. *Viruses* **10**(2018).
51. Manandhar, T., Ho, G.T., Pump, W.C., Blasczyk, R. & Bade-Doeding, C. Battle between Host Immune Cellular Responses and HCMV Immune Evasion. *Int J Mol Sci* **20**(2019).
52. Zeng, J. *et al.* Insights into the Transcriptome of Human Cytomegalovirus: A Comprehensive Review. *Viruses* **15**(2023).
53. Adamson, C.S. & Nevels, M.M. Bright and Early: Inhibiting Human Cytomegalovirus by Targeting Major Immediate-Early Gene Expression or Protein Function. *Viruses* **12**(2020).
54. Kim, D. *et al.* The architecture of SARS-CoV-2 transcriptome. *Cell* **181**, 914-921. e10 (2020).
55. Sunshine, S. *et al.* Systematic functional interrogation of SARS-CoV-2 host factors using Perturb-seq. *Nature communications* **14**, 6245 (2023).
56. Drayman, N., Patel, P., Vistain, L. & Tay, S. HSV-1 single-cell analysis reveals the activation of anti-viral and developmental programs in distinct sub-populations. *Elife* **8**(2019).
57. Labbe, K. & Saleh, M. Cell death in the host response to infection. *Cell Death Differ* **15**, 1339-49 (2008).
58. Augusto, L. *et al.* Toxoplasma gondii Co-opts the Unfolded Protein Response To Enhance Migration and Dissemination of Infected Host Cells. *mBio* **11**(2020).
59. Prasad, V. & Greber, U.F. The endoplasmic reticulum unfolded protein response–homeostasis, cell death and evolution in virus infections. *FEMS microbiology reviews* **45**, fuab016 (2021).
60. Saul, S. *et al.* Anticancer pan-ErbB inhibitors reduce inflammation and tissue injury and exert broad-spectrum antiviral effects. *J Clin Invest* **133**(2023).
61. Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**, 1103-1116 e20 (2020).
62. Joost, S. *et al.* The Molecular Anatomy of Mouse Skin during Hair Growth and Rest. *Cell Stem Cell* **26**, 441-457 e7 (2020).
63. Pethe, P., Noel, V.S. & Kale, V. Deterministic role of sonic hedgehog signalling pathway in specification of hemogenic versus endocardiogenic endothelium from differentiated human embryonic stem cells. *Cells Dev* **166**, 203685 (2021).
64. St-Jacques, B. *et al.* Sonic hedgehog signaling is essential for hair development. *Curr Biol* **8**, 1058-68 (1998).
65. Trevino, A.E. *et al.* Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. *Cell* **184**, 5053-5069 e23 (2021).
66. Wang, S. *et al.* The inconsistent regulation of HOXC13 on different keratins and the regulation mechanism on HOXC13 in cashmere goat (Capra hircus). *BMC Genomics* **19**, 630 (2018).

67. Peng, Y. *et al.* Detection of new pioneer transcription factors as cell-type-specific nucleosome binders. *Elife* **12**(2024).
68. Desanlis, I. *et al.* HOX13-dependent chromatin accessibility underlies the transition towards the digit development program. *Nat Commun* **11**, 2491 (2020).
69. Gillespie, D.T. The chemical Langevin Equation. *J. Chem. Phys.* **113**, 297-306 (2000).
70. Zwanzig, R. Ensemble method in the theory of irreversibility. *J. Chem. Phys.* **33**, 1338-1341 (1960).
71. Mori, H. Transport, collective motion, and Brownian motion. *Prog. Theor, Phys.* **33**, 423-455 (1965).
72. Xing, J. & Kim, K.S. Application of the projection operator formalism to non-Hamiltonian dynamics. *Journal of Chemical Physics* **134**(2011).
73. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587 e29 (2021).
74. Hastie, T. & Tibshirani, R. *Generalized additive models*, (Chapman and Hall, London ; New York, 1990).
75. Mitra, S. *et al.* Single-cell multi-ome regression models identify functional and disease-associated enhancers and enable chromatin potential analysis. *Nat Genet* **56**, 627-636 (2024).

Methods

Dynamical systems theory formulation of cellular state transitions

Assume that a cell state can be represented by the cell volume V (or cellular compartment size) and the copy number of $L >> 1$ pairs of gene products ($\boldsymbol{n}_m$, $\boldsymbol{n}_p$), where m and p designate mRNA and protein, and the bold fonts indicate vectors. For simplicity here we only consider m and p. It is straightforward to generalize to finer cell state specifications, for example, with distinction of nuclear and cytosol localizations, posttranslational states of proteins, other species such as microRNAs, epigenetic states, etc.

The temporal evolution of the cell state is described by a set of chemical master equations. When the copy numbers of molecular species are not too small, and the chemical reactions are not strongly coupled, Gillespie showed that the chemical master equations can be approximated by a set of chemical Langevin equations[69]. With extrinsic noises also included, we assume the ansatz that the dynamics of cell state is described by a set of generic stochastic differential equations,

$$\frac{d\boldsymbol{x}_m}{dt} = \boldsymbol{F}(\boldsymbol{x}_m, \boldsymbol{x}_p) + \zeta(\boldsymbol{x}_m, \boldsymbol{x}_p, t),$$

$$\frac{d\boldsymbol{x}_p}{dt} = \boldsymbol{G}(\boldsymbol{x}_m, \boldsymbol{x}_p) + \eta(\boldsymbol{x}_m, \boldsymbol{x}_p, t), \tag{4}$$

where the L-dimensional vectors $\boldsymbol{x}_m$ and $\boldsymbol{x}_p$ are cellular concentrations of m and p, and the ζ and η are taken as white noises with zero mean.

The low-dimensional manifold assumption is central to machine learning approaches on data analyses. From dynamical systems theory perspective, after a transient time a multi-dimensional dynamical system often converges to a low-dimensional slow manifold. In practice such property has been exploited with techniques such as quasi-steady-state approximation, quasi-equilibrium approximation. For a rigorous formulation, assume that one can identify a set of variables $(\boldsymbol{z}, \boldsymbol{Z})$ with $(2L\text{-}M)$ dimensional fast variables $\boldsymbol{z} = \boldsymbol{z}(\boldsymbol{x}_m, \boldsymbol{x}_p)$ and M-dimensional slower variables $\boldsymbol{Z} = \boldsymbol{Z}(\boldsymbol{x}_m, \boldsymbol{x}_p)$. Xing and Kim extended the celebrated Zwanzig-Mori projection[70,71] to a general dynamical system described by eq. 4[72]. The projection procedure results in a set of stochastic integral-differential equations of $\boldsymbol{Z}$ with colored noises, which are formally equivalent to eq. 4. Then if one assumes clear time scale separation between $\boldsymbol{z}$ and $\boldsymbol{Z}$, the equations reduce to a set of

Langevin equations with white noises, $\frac{d\boldsymbol{Z}}{dt} = \boldsymbol{A}\big(\boldsymbol{Z}(\boldsymbol{x}_m, \boldsymbol{x}_p)\big) + \eta(\boldsymbol{Z}, t)$, where $(\boldsymbol{Z}, t)$ are white noises with zero mean. Through ensemble averaging over the vicinity of a given point $\boldsymbol{Z}$, one has

$$\boldsymbol{v} \equiv < \frac{d\boldsymbol{Z}}{dt} >_{\boldsymbol{Z}} = \boldsymbol{A}\big(\boldsymbol{Z}(\boldsymbol{x}_m, \boldsymbol{x}_p)\big). \tag{5}$$

The equations define a M-dimensional manifold embedded in the $(\boldsymbol{x}_m, \boldsymbol{x}_p)$ space. A scRNA-seq data then measures the corresponding manifold projected to the transcriptomic subspace.

One should note that in practice the reported RNA velocity vector of a cell state i is typically obtained through averaging the raw velocity vectors of cell states within its neighborhood $\mathcal{N}_i$ on the manifold as a numerical approximation of the ensemble average, $\boldsymbol{v}_{\boldsymbol{x}i} = < \frac{d\boldsymbol{x}_m}{dt} > \approx \sum_{j \in \mathcal{N}_i} \boldsymbol{v}_{\boldsymbol{x}j}$. In this work, we used the k-nearest neighbor (kNN) algorithm to define the neighborhood in single modality datasets. For multi-omics data, the neighborhood of a cell state was defined using weighted nearest neighbors (WNN)[73] in the composite cell state space.

Mathematical foundation on applying GraphVelo to multi-omics datasets

Equation 3 in the main text applies to transformation between a manifold embedded in a state space and in a subspace. According to the Whitney Embedding theorem[18], any smooth real M-dimensional manifold can be embedded in a $2M$-dimensional real space provided that $M > 0$. Consider a full set of genes versus a subset in a scRNAseq dataset, or a combined scRNAseq/scATACseq multi-omics dataset versus the scRNAseq subset. Assume that the full cell state space has a dimensionality N, while a single cell data manifold is typically low-dimensional with $M << N$. Then the Whitney Embedding theorem[18] suggests that with proper choice of the subset the manifolds in the full space and the subspace are homeomorphic or at least piece-wise homeomorphic (Fig. 1b, Supplemental Text IV), i.e., a one-to-one mapping exists between the two. Then applying eq. 3 allows one to infer the velocity vectors for the full-space representation from those of the subspace.

Denoise velocity vectors in the space of principal components (PCs). Learning coefficients $\boldsymbol{\phi}_i$ in the gene space directly often fails due to thousands of gene profiles. To avoid the curse of high dimensionality and learn parameters in a compact manifold, we designed a procedure to denoise the velocities in a reduced PCA space. Specifically, we extrapolated the cell state i in the original space using the infinitesimal propagation operator to extrapolate the future state:

$$\boldsymbol{x}_i' = \boldsymbol{x}_i + \boldsymbol{v}_i \cdot dt_i$$

Moreover, we estimated an optimal step size *dt* based on the local density to guarantee the cell states are bound to the manifold:

$$dt_i = \text{median}\left(\frac{\frac{1}{k}\sum_{j=1}^{k}\|\boldsymbol{\delta}_{ij}\|}{\|\boldsymbol{v}_i\|}\right)$$

After utilizing the cell-dependent t_i to forcing the predictions inhabiting regions of the phenotypical manifold, we applied dimension reduction to project both current and future status from the gene space to the PCA space through linear transformation. Then we obtained the projected velocity vectors as:

$$\boldsymbol{v}_i^{\text{PCA}} = \frac{\boldsymbol{x}_i' Q - \boldsymbol{x}_i Q}{dt_i}$$

where Q is the PC loading matrix estimated using x that serves as the coordinate transformation matrix.

Manifold-consistent kinetic genes. The kinetic assumptions between nascent and mature RNA fail when the underlying parameters shift along the developmental trajectory[11], which leads to transcription burst and rapid degradation in the phase portraits (Fig. 3a). An internal clock exists during cell proliferation and differentiation. Current methods rely on different criteria to select confident estimated velocity genes (see Supplemental Text V for detailed discussion). Here, we presume that the velocity of robustly estimated genes should be consistent with the (pseudo)time derivative estimated under the manifold assumption. We can utilize any available *t*s inferred from data manifold by either pseudotime, velocity latent time, or lineage tracing to approximate the temporal information and use *k* nearest neighbors to define the locally linear plane. After ordering cells within the local Euclidean space, we calculate the MacK score for any gene g as an indicator of whether the sign of estimated velocity agrees with the dynamic cascades within manifold,

$$\text{MacK score} = \frac{1}{n}\sum_{j\in\mathcal{N}(i)} \mathbb{I}\left(sgn\left(\frac{\Delta x_{ij}(g)}{\Delta t_{ij}}\right) = sgn(v_i(g))\right)$$

Where $\mathcal{N}_i$ indicates the neighbor points of cell *i*, $\mathbb{I}$ represents the indicator function and *sgn* returns the sign of the values. $\Delta x_{ij}(g), v_i(g)$ are the difference in abundance of gene g between cell *i* and

j, and the velocity of gene g in cell i, respectively. We parallelize the calculation to scale efficiently with the number of genes, which is important due to the number of highly variable genes. We want to point out that one can use methods other than the MacK score to identify genes with reliable velocity estimations.

Dynamo criteria. Dynamo[7] offers a correction strategy by removing genes with low gene-wise confidence in the phase plane. This allows us to identify genes that appear in incorrect phase portrait positions and contribute to erroneous flow directions (illustrated in Fig3. a). To filter out genes with misleading dynamic patterns, one can supply the established lineage hierarchy information to the *dyn.tl.confident_cell_velocities* function in dynamo. This function scores each gene based on the agreement of its behavior in the splicing phase diagram with the input lineage hierarchy priors.

Post-GraphVelo analyses

Reconstruction of extended dynamo vector field from multi-omics data

Dynamo is a general framework of reconstructing dynamical models from scRNAseq data, and it is straightforward to generalize to multi-modal data. The framework is based on specific realizations of eq. 5, $\boldsymbol{v} \equiv < \frac{d\boldsymbol{x}}{dt} >_{\boldsymbol{x}} = \boldsymbol{A}(\boldsymbol{x})$, with state vector $\boldsymbol{x}$ being the transcript concentrations for scRNAseq data, and combined transcript concentrations and continuous quantification of locus-specific chromatin open-close state for multi-omics scRNAseq/scATACseq data. The variables $\boldsymbol{x}$ can be defined in various representations, e.g., the original gene space, principal component subspace, latent space defined by variational autoencoder, etc. With GraphVelo it is straightforward to transform $\boldsymbol{v}_{\boldsymbol{x}}$ between different representations.

The continuous vector field functions $\boldsymbol{A}(\boldsymbol{x})$ contain quantitative regulation relations between genes that are learned from single cell data points of $(\boldsymbol{x}, \boldsymbol{v}_{\boldsymbol{x}})$. Various algorithms can be used to learn the *analytical* forms of $\boldsymbol{A}(\boldsymbol{x})$. The original dynamo paper illustrated a Reproducing Kernel Hilbert Space (RKHS) representation method. The method expresses $\boldsymbol{A}(\boldsymbol{x})$ as a linear combination of pre-selected basis functions, $\boldsymbol{v} = \boldsymbol{A}(\boldsymbol{x}) = \sum_{\alpha} \boldsymbol{c}_{\boldsymbol{\alpha}} \boldsymbol{\Gamma}_{\alpha}(\boldsymbol{x})$, similar to the more familiar Taylor expansion that uses a linear combination of polynomial functions to represent a continuous analytical function. It should be noted that the basis functions and so $\boldsymbol{A}(\boldsymbol{x})$ are generally *nonlinear* functions of $\boldsymbol{x}$. Following Qiu et al.[7], we chose Gaussian functions centered at selected reference points $\tilde{\boldsymbol{x}}_{\alpha}$,

$\boldsymbol{\Gamma}_a(\boldsymbol{x}, \widetilde{\boldsymbol{x}}_\alpha) = e^{-2w\|\boldsymbol{x}-\widetilde{\boldsymbol{x}}_\alpha\|^2}$, with default parameter value of w in the package dynamo. Then we determined the coefficient vectors $\boldsymbol{c}_{\boldsymbol{\alpha}}$through minimizing the loss function $\Phi(\boldsymbol{c}_1, \boldsymbol{c}_2, \dots, \boldsymbol{c}_{\mathrm{m}}) = \sum_{i=1}^{n}\|\boldsymbol{v}_i - \sum_\alpha \Gamma(\boldsymbol{x}, \widetilde{\boldsymbol{x}}_\alpha)\boldsymbol{c}_\alpha\|^2 + \frac{\lambda}{2}\sum_{\alpha=1}^{m}\sum_{\beta=1}^{m}\boldsymbol{c}_\alpha^T\Gamma(\widetilde{\boldsymbol{x}}_\alpha, \widetilde{\boldsymbol{x}}_\beta)\boldsymbol{c}_\beta$, where the first sum was over all the data points, and the second term was Tikhonov regularization weighted by λ. The superscript T means matrix transpose. One can also use neural networks, e.g., variational autoencoders, to learn an optimal set of basis functions, and other algorithms such as neural ODE to learn $\boldsymbol{B}(\boldsymbol{x})$. The difference is merely algorithmic under the same framework of dynamical systems theories.

Jacobian analyses and reconstruction of effective dose-response curves of gene regulation. The extended dynamo vector field contains generally nonlinear relationship about regulation between genes, and between genes and other modalities (e.g., chromatin open/close conformations). Several posterior interpretation methods exist to analyze the vector field. Below we will describe two of them.

With the analytical form of **F**(x), one can calculate efficiently the Jacobian field **J** at any cell state **x**. Each element of **J**, ($J_{ij} = \partial F_i/\partial x_j$), can be understood as an in-silico perturbation experiment on how upregulating gene *j* affects the transcription rate of gene *i,* with all other gene expression levels kept constant at state **x**. For example, a positive value of J_{ij} indicates that at **x** further increasing the expression of gene *j* causes increase of the transcription rate of gene *i*. Note that the sign and value of a Jacobian element alone does not unambiguously reflect the nature of the regulation. A close-to-zero Jacobian element can be associated with either no regulation or the regulator is at a saturating concentration of regulation. The regulation relation can be direct, or indirect through intermediate molecular species not implicitly treated as variables of the vector field function.

Complementary to the local Jacobian analysis is to reconstruct effective dose-response (D-R) curves of a regulator-target gene pair. The curve reveals the rate of change of one quantity, e.g., the transcription rate or the chromatin open/close dynamics of the target gene, as a function of the value of a regulator, e.g., the mRNA level of a regulator or the chromatin open/close status of a specific genomic region. Note that the D-R curve is generally a multi-variate function, we designed a procedure to reconstruct an effective one-variate function[7]. One can genetically write the regulation on quantity x_i as two terms with and without dependence on variable x_j,

$$\frac{dx_i}{dt} = F_i(x_1, \dots, x_N) = F_i^1(x_1, \dots, x_N) + F_i^2(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_N).$$

Notice that F_i^2 is not a function of gene j. First, we calculated the Jacobian element $\frac{\partial F_i}{\partial x_j}$ for each measured cell state. Note $\frac{\partial F_i}{\partial x_j} = \frac{\partial F_i^1}{\partial x_j}$, so the background variation from F_i^2 due to effects of other genes has been numerically removed. Then from the histogram of $\frac{\partial F_i}{\partial x_j}$ versus x_j, we binned $\frac{\partial F_i}{\partial x_j}$ and calculated $< \frac{\partial F_i}{\partial x_j} >_{x_i}$ for each bin. Next, we performed numerical integration to obtain $\overline{F_i^1}(x_i) = F_{i0}^1 + \int_0^{x_i} < \frac{\partial F_i}{\partial x_j} >_{x_i} dx_i \approx F_{i0}^1 + \sum_\alpha < \frac{\partial F_i}{\partial x_j} >_{x_{i\alpha}} \Delta x_{i\alpha}$. In practice, if $\frac{\partial F_i}{\partial x_j}$ shows large variance within each bin of x_j, it may imply that other factors affect the D-R curve. For example, the regulation of x_j on x_i may even be opposite at the presence or absence of a specific cofactor. In this case, one should first cluster cells, e.g., grouping cells based on whether an identified cofactor reaches a threshold value, then perform the D-R curve reconstruction on individual clusters.

Markovian analyses. Marius et al.[16] have developed a framework named CellRank to study cellular dynamics based on Markov chain formulation. We use CellRank to identify cell state transitions using velocity kernel and identify terminal states within datasets by GPCCA function module. In addition, CellRank pseudotime kernel[8] is used for methods comparison in real datasets.

Functional form vector field and differential geometry analyses. Dynamo first proposed to learn nonlinear functional form of RNA velocity vector field, providing a physical-informed framework which integrates mechanism modeling and single cell data analyses. We use dynamo to learn continuous vector filed and perform differential geometry analyses such as gene acceleration, vector field-based pseudotime, LAP, Jacobian analyses and in silico perturbation.

Pseudotemporal orderings. GraphVelo itself does not compute an ordering index of cells as we are seeking for a more quantitative method to infer RNA velocity. With an accurate RNA velocity as input, we can approximate the vector field precisely. Thus, we use the scalar potential estimated from the functional form vector field with Hodge decomposition as a proxy of time, which is implemented by dynamo package.

Cross boundary correctness scoring. While the GraphVelo framework is designed to quantify velocity vectors across different representations, known transitions between coarse cell states—such as cell types or cell cycle phases—can be used to evaluate the correctness of velocity directions. Suppose there are two cell populations A and B, with A a progenitor state of B. One can define the set of boundary cells between A and B as

$$\mathcal{C}_{\mathrm{A}\to\mathrm{B}} = \{c \in \mathcal{C}_{\mathrm{A}} | \exists c' \in \mathcal{C}_{\mathrm{B}} \cap \mathcal{N}_c\},$$

where $\mathcal{C}_{\mathrm{A}}$ or $\mathcal{C}_{\mathrm{B}}$ denote the sets of cells in state A or state B, $\mathcal{N}_c$ indicates the kNN of cell c. The CBC score is then defined as

$$\text{CBC score} = \frac{1}{n}\sum_c \frac{1}{\#c' \in \mathcal{C}_{\mathrm{B}} \cap \mathcal{N}_c} \sum_{c' \in \mathcal{C}_{\mathrm{B}} \cap \mathcal{N}_c} \frac{\boldsymbol{v}_c \cdot (\boldsymbol{x}_{c'} - \boldsymbol{x}_c)}{|\boldsymbol{v}_c| \cdot |\boldsymbol{x}_{c'} - \boldsymbol{x}_c|},$$

where $\#c' \in \mathcal{C}_{\mathrm{B}} \cap \mathcal{N}_c$ is the number of cells in state B which is also the kNN of cell c. While the $(\boldsymbol{x}_c, \boldsymbol{v}_c)$ can be represented in different basis (raw count, PCA or UMAP), we computed the CBC score in the original count space to ensure that all genes contribute to the velocity estimation. We deliberately avoided using 2D embeddings like UMAP for this purpose, as such visualizations may distort the true geometric relationships in the high-dimensional space and could lead to misleading interpretations.

Approximating smooth velocity trends with generalized addictive model. The variation of transcription rates contains the high order dynamic information of the cell system. To model the dynamic patterns of RNA velocity along transition path from noisy data, we refine the velocity vectors by local geometry via TSP and further fit GAM to velocity value of each gene that has been refined by GraphVelo. For any gene g, we model the velocity trend for the temporal variable t via

$$v_{gi} = \beta_0 + f(t_i)$$

Where v_{gi} indicates the velocity of gene g in cell i, f is built using penalized B-splines which allow us to automatically model non-linear mapping while maintaining additivity[74]. To visualize the velocity trends, we select 100 equally spaced testing points along transition path and predict gene expression at each of them using the fitted model. The estimated velocity trends can be treated as smoothed time series for further analyses.

Clustering velocity trends along infection trajectory. With manifold-constrained velocity estimated by GraphVelo, we are able to cluster genes into different functional modules which are

involved in the same regulatory circuit. We recover transcription variation of both host and virus factors along infection trajectory by fitting GAMs in temporal indicator, the percentage of viral RNA. Next, we select 100 equally spaced time points and generate the GAM-smoothed velocity trends. We compute a kNN graph and cluster the kNN graph using the leiden algorithm. We used k = 15 for the velocity-trend kNN graph and the leiden algorithm with resolution parameter set to 0.3 to avoid over clustering the trends. For each recovered cluster, we compute its mean and standard deviation (pointwise, for all generated points that were used for smoothing) and visualize the smoothed trends per cluster.

Characterizing decoupling genes based on dynamic time warping distance between multi-modality velocities. We perform the DTW distance calculation by dtaidistance package. To eliminate the influence of scale in different modality, we maximum-normalize the chromatin/RNA velocity to the same range of [0, 1]. Then we fit the velocity trends of both modalities along vector field-based pseudotime to yield the smoothed velocity trends. We calculate the DTW distance between velocity trends per gene. To distinguish the decoupling genes based on their multi-modality velocities, we rank the genes based on the DTW distance and identify elbow point as cutoff. Genes with distance metric larger than the cutoff are characterized as decoupling genes and used for visualization and functional analysis.

Synthetic datasets

Generating two genes bifurcation process and mapping it to 3d. The bifurcation data (n=2,000 cells) for the toggle-switch system is simulated using the Gillespie algorithm. We use activation and inhibition Hill functions to model the induction and suppression effects between the two genes:

$$\dot{x} = \frac{a_1 x^n}{S_1^n + x^n} + \frac{b_1 K_1^n}{K_1^n + y^n} - \gamma_1 x,$$

$$\dot{y} = \frac{a_2 y^n}{S_2^n + y^n} + \frac{b_2 K_2^n}{K_2^n + x^n} - \gamma_2 y,$$

We use the simulation backend implemented by dynamo with default parameters except the timescale (reset τ =1) to generate the bifurcating process. We then map the synthetic dataset onto a sphere (radius r=70) and yield the variable z as:

$$z = \sqrt{max(r^2 - x^2 - y^2, 0)}$$

Then we are able to calculate the correctly-scaled 3d vectors by infinitesimal propagation operator with sufficient small step size (*dt*=1 in our case):

$$\dot{z} = \lim_{\Delta t \to 0} \frac{\sqrt{\max(r^2 - (x + \dot{x}\Delta t)^2 - (y + \dot{y}\Delta t)^2, 0)} - z}{\Delta t}$$

Generating scRNA-seq synthetic data with dyngen. To generate high-dimensional single-cell transcriptomic data in silico, we use a multi-modal simulation engine, dyngen, to account for different developmental topologies. We constructe module networks to represent regulatory cascades and feedback loops driving progressive changes in gene expression and influencing cell fate decisions. We generate three datasets with 1,000 cells and 100 genes using the linear, cyclic, and bifurcating loop backbones provided by dyngen, with all other parameters set to default values. These datasets include simulated nascent and mature mRNA counts along with ground-truth RNA velocities and known manifold structure.

Simulation of rapid degradation and transcription burst events on phase portrait. As for genes with variable degradation rate, we present a minimal regulatory network with linear model in which an external signal both inhibits transcription and promotes microRNA (miRNA). The miRNA exerts a linear influence on the degradation rate of mRNA. We have miRNA's velocity as

$$\frac{dm}{dt} = \alpha_m - \gamma_m m$$

The nascent gene transcription rate and mature mRNA's degradation rate would change to

$$\alpha = \alpha_0 - k_\alpha t$$

$$\gamma = \gamma_0 + k_\gamma m$$

where α_0 and γ_0 represent the constant transcription rate and degradation rate without the effect of miRNA, k_α and k_γ represent the magnitude of influence from miRNA, *t* indicates the simulation time to mimic the cell-context change along trajectory.

$$\frac{du}{dt} = \alpha - \beta u \ = \ \alpha_0 - k_\alpha m \ - \beta u$$

$$\frac{ds}{dt} = \beta u - \gamma s = \beta u - (\gamma_0 + k_\gamma m)s$$

We set the initial condition to a steady state with $u_0 = \alpha_0/\beta$ and $s_0 = \alpha_0/\gamma_0$, while the miRNA abundance $m_0 = 0$. We simulate u_t and s_t as the microRNA signal m_t gradually increases. The

892 aim is to evaluate whether the estimated RNA velocity consistently aligns in sign with the ground
893 truth RNA velocity.

894 To generate genes with transcription burst phase portrait, we set the initial condition to $u_0 = 0$,
895 $s_0 = 0$, together with γ as constant and α promotes to $\alpha' = 3\alpha$ when simulation reaches specific
896 time.

897 **Processing of sequencing data**

898 All sequencing data in this study are downloaded publicly (see details in the 'Data availability'
899 section). Though the number of confident-estimated gene sets differed case by case, GraphVelo
900 predicted the velocity of all highly variable genes and used them for downstream calculation such
901 as CBC score, CellRank velocity kernel and dynamo vector field learning. We parallelized the
902 TSP optimization to scale efficiently and set the hyperparameters in loss with a = 1, b = 10 and
903 λ = 1 as default in all studies (Extended Data Fig 16).

904 **Analyses of scRNA-seq datasets.** For the erythroid lineage of the mouse gastrulation, we follow
905 the standard data pre-processing procedures implemented by scVelo and select 9,815 cells and
906 2,000 highly variable genes to construct the k-nearest neighbor graph using 30 nearest neighbors
907 for downstream calculation.

908 For the human bone marrow dataset, we follow the standard data pre-processing procedures
909 implemented by scVelo and selected 5,780 cells and 2,000 highly variable genes to construct the
910 *k*-nearest neighbor graph using 30 nearest neighbors for downstream calculation. To estimate the
911 variation of degradation rate γ along the differentiation lineages, we divide the cells into five
912 discrete time bins based on precomputed Palantir pseudotime. We then estimate cell-specific
913 degradation rates and visualized their distribution shifts along the hematopoiesis trajectory.

914 **Analyses of HCMV dataset.** We sample the cells from donor 1 to eliminate sample-specific
915 variation and further filter out cells lacking immediate early *UL123* gene expression to focus on
916 the viral infection trajectory. We adapt monocle preprocessing recipe implemented by dynamo and
917 yield 1,454 cells and 2,000 highly variable genes for further analyses. The number of nearest
918 neighbors were set to 30 as default.

919 We collect the pathway-related genes from MSigDB and perform the Jacobian analyses
920 implemented by dynamo, using viral genes as regulators and host genes as effectors. We rank the
921 regulation relationships based on the collections of Jacobian elements. We pick top 50 inhibited
922 effectors of viral genome and select the common set between pathway genes and all the effectors
923 for visualization.

924 In silico knock out experiments are performed via dynamo. We suppressed every single virus
925 factors per time using *dynamo.pd.perturbation* function and calculated the change of total viral
926 gene expression after perturbation.

927 **Analyses of SARS-CoV-2 dataset.** We subsampled the cells based on their infection status. We
928 adapted the monocle preprocessing pipeline implemented in dynamo and excluded the UMI reads
929 from lenti-virus. We obtained 5,001 infected cells and 2,000 highly variable genes for downstream
930 analyses. We used both highly variable genes in host and virus genes to perform PCA and
931 calculated the nearest neighbors with k = 30. We used the apoptosis-related genes collected by
932 KEGG to calculate apoptosis activity score of each cell using *dynnamo.tl.score_cells* function.

933 **Analyses of multi-modality datasets.**

934 **RNA velocity estimation for multi-modality datasets.** In traditional scRNA-seq datasets, RNA
935 velocity methods use smoothed spliced and unspliced RNA counts through nearest-neighbor
936 pooling, based on the PCA space computed from transcripts alone. However, this approach is not
937 suitable for multimodal scenarios, as it overlooks hidden variables by relying on a single modality.
938 To construct a consistent manifold combining information from multi-modal genomics data, we
939 utilized WNN as implemented in MultiVelo[13]. The WNN algorithm combines low-dimensional
940 representations from RNA and ATAC omics data. Specifically, we use PCA results from scRNA-
941 seq data and latent semantic indexing (LSI) from scATAC-seq as inputs. The nearest neighbors
942 identified by WNN were then used to calculate the first moment, reducing noise in separate
943 modalities and approximating a unified manifold for GraphVelo.

944 **Mouse skin dataset preprocessing.** The preprocessed SHARE-seq mouse skin dataset[61] is
945 adopted directly from MultiVelo data resources. All the procedures are consistent with MultiVelo
946 except we get the LSI representation processed by SCARlink[75]. We construct the WNN graph

using 50 nearest neighbors for downstream calculation. We run scVelo with 'stochastic' mode to estimate the RNA velocity based on the WNN graph as we discussed above.

Human cortex dataset preprocessing. The preprocessed human cerebral cortex data is adopted directly from MultiVelo data resources. All the procedures are consistent with MultiVelo except we get the LSI representation processed by SCARlink. We construct the WNN graph using 50 nearest neighbors for downstream calculation. We run scVelo with 'stochastic' mode to estimate the RNA velocity based on the WNN graph.

Data availability

All the sequencing raw data are publicly accessible. The A549 dataset can be accessed via https://figshare.com/ndownloader/files/53666738. The FUCCI cell cycle data can be downloaded from https://figshare.com/ndownloader/files/53705057. The mouse gastrulation subset to erythroid lineage can be extracted using scVelo's CLI: *scvelo.datasets.gastrulation_erythroid()* or from the original work under accession number E-MTAB-6967 of ArrayExpress. The human bone marrow can be extracted using scVelo's CLI: *scvelo.datasets.bonemarrow()* or through the Human Cell Atlas data portal at https://data.humancellatlas.org/explore/projects/091cf39b-01bc-42e5-9437-f419a66c8a45. The original HCMV infected moDC data can be accessed via Zenodo (https://zenodo.org/records/10404879) and the processed data can be downloaded from https://figshare.com/ndownloader/files/53666756. The SARS-CoV-2 data can be accessed via https://figshare.com/ndownloader/files/53666588. The preprocessed mouse skin development dataset can be accessed via https://figshare.com/articles/dataset/Mouse_Hair_Follicle_RNA_Data/22575307 and https://figshare.com/articles/dataset/Mouse_hair_follicle_ATAC_data/22575313. The preprocessed human cortex dataset can be downloaded from https://figshare.com/articles/dataset/Developing_Human_Cortex_RNA_Data/22575376 and https://figshare.com/articles/dataset/Developing_Human_Cortex_ATAC_Data/22575370.

Code availability

Python package GraphVelo can be accessed from https://github.com/xing-lab-pitt/GraphVelo. Reproducibility and tutorials can be found in https://github.com/xing-lab-pitt/GraphVelo/tree/main/notebook and https://graphvelo.readthedocs.io/en/latest/.

References

54. Gillespie, D.T. The chemical Langevin Equation. *J. Chem. Phys.* **113**, 297-306 (2000).
55. Zwanzig, R. Ensemble method in the theory of irreversibility. *J. Chem. Phys.* **33**, 1338-1341 (1960).
56. Mori, H. Transport, collective motion, and Brownian motion. *Prog. Theor, Phys.* **33**, 423-455 (1965).
57. Xing, J. & Kim, K.S. Application of the projection operator formalism to non-Hamiltonian dynamics. *Journal of Chemical Physics* **134**(2011).
58. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587 e29 (2021).
59. Hastie, T. & Tibshirani, R. *Generalized additive models*, (Chapman and Hall, London ; New York, 1990).
60. Mitra, S. *et al.* Single-cell multi-ome regression models identify functional and disease-associated enhancers and enable chromatin potential analysis. *Nat Genet* **56**, 627-636 (2024).

Author contributions

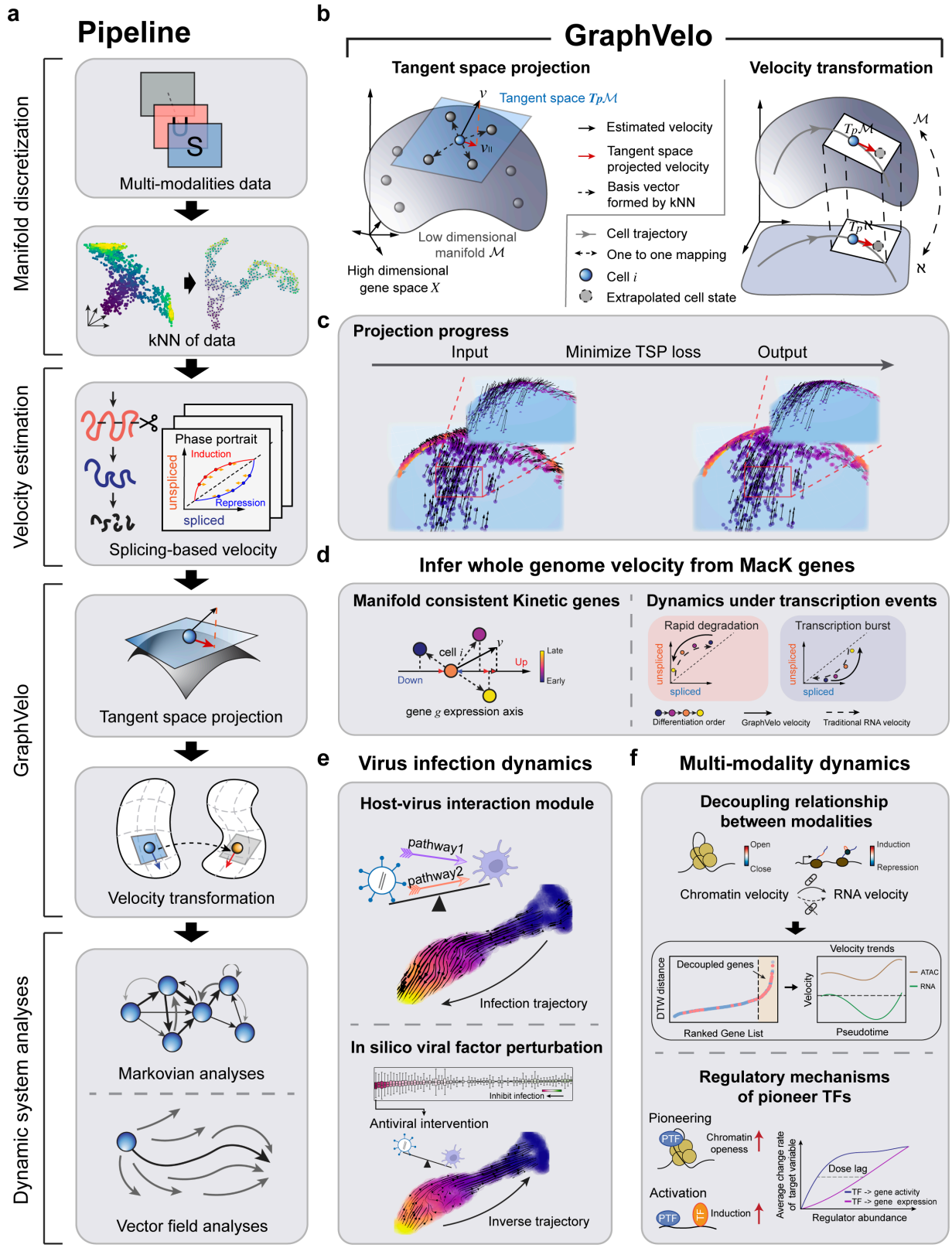
JX conceived and formulated the theoretical framework, YC, and JG performed tests on simulated data, YC and JX analyzed scRNA-seq datasets, YC, YZ, JG, and KN developed the package GraphVelo, YC and JX wrote the draft, and all authors edited the manuscript.

Acknowledgments

This work was partially supported by National Institute of General Medical Sciences (1R01GM148525 to JX and 1R01GM139297 to IB), and National Science Foundation (2325149) to JX.

This work was supported by the National Key Research and Development Program of China (2023YFE0112300); National Natural Sciences Foundation of China (32261133526; 32270709; 32070677); the 151 talent project of Zhejiang Province (first level), the Science and Technology Innovation Leading Scientist (2022R52035)

Figures:


a
Pipeline
Manifold discretization
Multi-modalities data
kNN of data
Velocity estimation
Phase portrait
Induction
Repression
unspliced
spliced
Splicing-based velocity
GraphVelo
Tangent space projection
Velocity transformation
Dynamic system analyses
Markovian analyses
Vector field analyses
b
GraphVelo
Tangent space projection
Tangent space $T_p\mathcal{M}$
Estimated velocity
Tangent space projected velocity
Basis vector formed by kNN
Cell trajectory
One to one mapping
Cell i
Extrapolated cell state
Low dimensional manifold $\mathcal{M}$
High dimensional gene space X
Velocity transformation
c
Projection progress
Input
Minimize TSP loss
Output
d
Infer whole genome velocity from MacK genes
Manifold consistent Kinetic genes
cell i
Up
Down
Late
Early
gene g expression axis
Dynamics under transcription events
Rapid degradation
Transcription burst
Differentiation order
GraphVelo velocity
Traditional RNA velocity
e
Virus infection dynamics
Host-virus interaction module
pathway1
pathway2
Infection trajectory
In silico viral factor perturbation
Inhibit infection
Antiviral intervention
Inverse trajectory
f
Multi-modality dynamics
Decoupling relationship between modalities
Open
Close
Induction
Repression
Chromatin velocity
RNA velocity
Velocity trends
Decoupled genes
DTW distance
Ranked Gene List
Velocity
Pseudotime
ATAC
RNA
Regulatory mechanisms of pioneer TFs
Pioneering
PTF
Chromatin openess
Activation
TF
Induction
Average change rate of target variable
Dose lag
TF -> gene activity
TF -> gene expression
Regulator abundance

1002 **Figure 1. Refining RNA velocity by tangent space projection and transforming between**
1003 **representations using GraphVelo.**

1004 **(a)** Workflow of RNA velocity-based analyses incoporating GraphVelo. Note GraphVelo takes
1005 any form of RNA velocity (i.e., not just splicing-based velocity) as input, and the kNN
1006 neighborhood is defined in the full state space (e.g., by both scRNAseq and scATACseq in multi-
1007 omics data).

1008 **(b)** Schematic of tangent space projection and velocity transformation between homeomophic
1009 manifolds. *Left*: RNA velocity vectors are projected onto the tangent space defined by the
1010 discretized local manifold of neighborhood cell samples. *Right*: GraphVelo allows for
1011 transformation of velocity vectors from a manifold embeded in a higher dimensional space ($\mathcal{M}$)
1012 to that in a lower dimensional space ($\aleph$), and *vice versa*.

1013 **(c)** The process of minimizing the loss function of tangent space projection. Noisy velocity vectors
1014 (*left*) generated by adding random components orthogonal to those sampled from an analytical 2D
1015 manifold were projected back onto the 2D manifold, resulting in smooth velocity vectors that lie
1016 in the tangent space (*right*).

1017 **(d)** GraphVelo allows whole genome velocity inference based on the robustly estimated MacK
1018 genes (see also Figure 3). Velocities of genes undergoing variable kinetic rates, such as rapid
1019 degradation or transcription burst, are difficult to be correctly inferred by other methods, but can
1020 be inferred robustly with GraphVelo.

1021 **(e)** Virus infection dynamics and underlying host-virus interaction mechanisms uncovered by
1022 GraphVelo (see also Figure 4). *Upper*: pathways involved in host-virus interactions were identified
1023 using GraphVelo. *Lower*: GraphVelo predicted reversed trajectory of viral infection in response to
1024 *in silico* perturbations of viral factors.

1025 **(f)** GraphVelo provides a consistent view of epigenetic and transcription dynamics (see also Figure
1026 5). *Upper*: GraphVelo analyses on multio-mics data revealed that most cell-cycle dependent genes
1027 showed decoupling between transcription dynamics and chromatin accessibility change dynamics.
1028 *Lower*: Effective dose-response curves reconstructed from multi-omics data revealed pioneer
1029 transcription factors increased chromatin accessibility then transcription of targe genes.

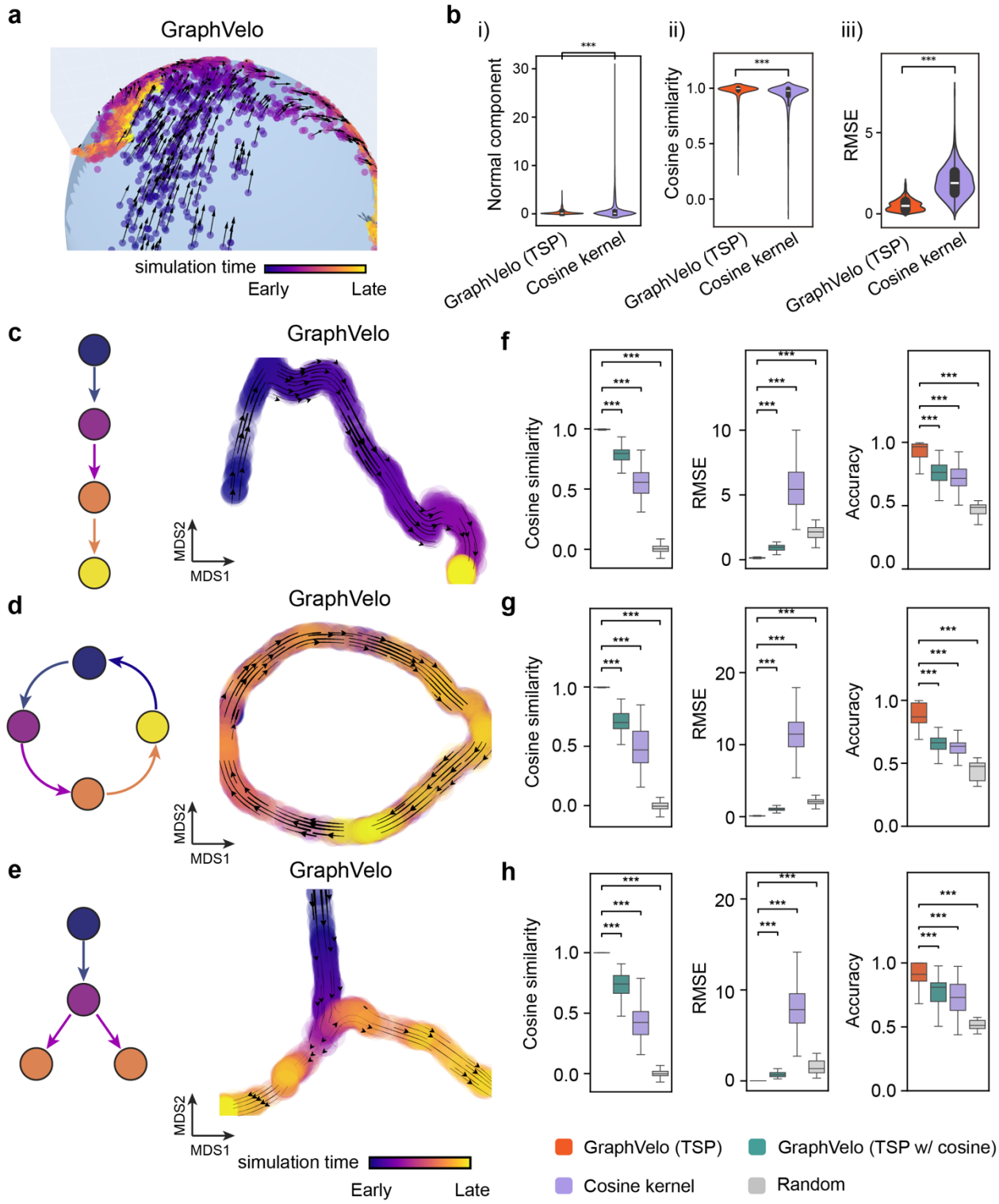


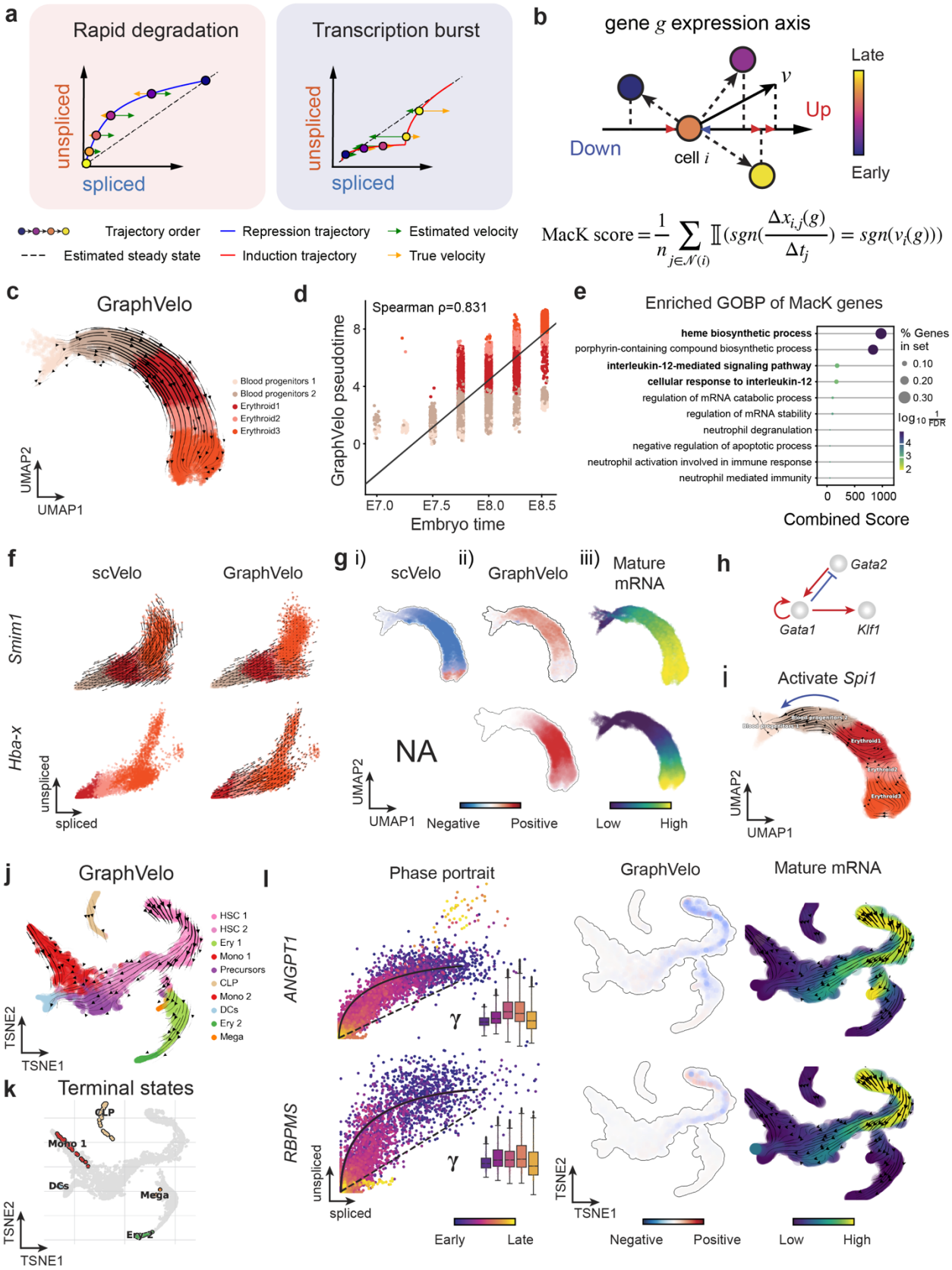
Figure 2. Testing graphVelo on simulated datasets.

(a) Velocity vectors of an analytical three variables bifurcating vector field constrained to a spherical surface. The data points were colored by simulation time.

(b) Violinplots of: i) normal component of velocity vectors, ii) cosine similarity and iii) root mean square error (RMSE) between ground truth and velocity vectors projected by GraphVelo and cosine kernel, respectively.

(c-e) Simulation of scRNA-seq data using dyngen under linear, cycling, and bifurcating differentiation models (*left*), and velocity fields projected on multidimensional scaling (MDS) coodinates (*right*) using GraphVelo-corrected velocities, respectively. Each simulation consists of 1,000 cell states and 100 genes. The cells in different states were colored by their simulation time along trajectory.

(f-h) Boxplots of cosine similarity, , and accuracy between the ground truth velocity vectors and dyngen simulated velocities after projection using GraphVelo TSP loss without cosine regularization, GraphVelo TSP loss with cosine regularization, cosine kernel, and random predictor, respectively.


a
Rapid degradation
unspliced
spliced
Transcription burst
unspliced
spliced
Trajectory order
Estimated steady state
Repression trajectory
Induction trajectory
Estimated velocity
True velocity
b
gene g expression axis
Late
Early
Up
Down
cell i
v
MacK score = 1/n Σ_{j∈N(i)} 𝕀(sgn(Δx_{i,j}(g)/Δt_j) = sgn(v_i(g)))
c
GraphVelo
Blood progenitors 1
Blood progenitors 2
Erythroid1
Erythroid2
Erythroid3
UMAP2
UMAP1
d
Spearman ρ=0.831
GraphVelo pseudotime
Embryo time
E7.0
E7.5
E8.0
E8.5
e
Enriched GOBP of MacK genes
heme biosynthetic process
porphyrin-containing compound biosynthetic process
interleukin-12-mediated signaling pathway
cellular response to interleukin-12
regulation of mRNA catabolic process
regulation of mRNA stability
neutrophil degranulation
negative regulation of apoptotic process
neutrophil activation involved in immune response
neutrophil mediated immunity
% Genes in set
0.10
0.20
0.30
log10 1/FDR
Combined Score
f
scVelo
GraphVelo
Smim1
Hba-x
unspliced
spliced
g
i) scVelo
ii) GraphVelo
iii) Mature mRNA
NA
UMAP2
UMAP1
Negative
Positive
Low
High
h
Gata2
Gata1
Klf1
i
Activate Spi1
UMAP2
UMAP1
j
GraphVelo
HSC 1
HSC 2
Ery 1
Mono 1
Precursors
CLP
Mono 2
DCs
Ery 2
Mega
TSNE2
TSNE1
k
Terminal states
CLP
Mono 1
DCs
Mega
Ery 2
TSNE2
TSNE1
l
Phase portrait
GraphVelo
Mature mRNA
ANGPT1
RBPMS
γ
unspliced
spliced
Early
Late
TSNE2
TSNE1
Negative
Positive
Low
High

1048

1049 **Figure 3. Delineating transcriptome-wise progression with manifold-consistent kinetic genes**
1050 **using GraphVelo**

1051 **(a)** Schematitc of transcritional events mislead RNA velocity estimation in the phase portrait by
1052 standard approaches. *Left*: for genes exhibiting rapid degradation, the cells appear above the steady
1053 state line on the phase portrait, whereas the true velocity is negative. *Right*: For genes exhibiting
1054 transcription burst, the transcription rate abruptly increases at intermediate states, leading to a
1055 steady state line whose slope is overestimated.

1056 **(b)** Schematic of manifold-consistent score calculation for robustly estimated velocity genes.

1057 **(c)** The projected velocity field from GraphVelo are consistent with the erythroid differentiation
1058 by using all highly variable genes.

1059 **(d)** The correlation between GraphVelo vector field-based pseudotime and embryo time for
1060 erythroid lineage cells. Spearman correlation coefficients are shown.

1061 **(e)** GO enrichment analyses of top ranked MacK genes.

1062 **(f)** The phase portrait of two transcription burst genes (*Smim1*, *Hba-x*).

1063 **(g)** Scatter plots of: i) velocities estimated by scVelo, ii) refined velocities by GraphVelo, and iii)
1064 mature mRNA expression of transcription burst genes (*Smim1*, *Hba-x*). Cells were colored by
1065 corresponding velocity, and mature mRNA abundance, respectively, and visualized on the UMAP
1066 representation.

1067 **(h)** Gene regulatory cascade unraveled by GraphVelo-based vector field analyses that drives cell
1068 lineage commitment.

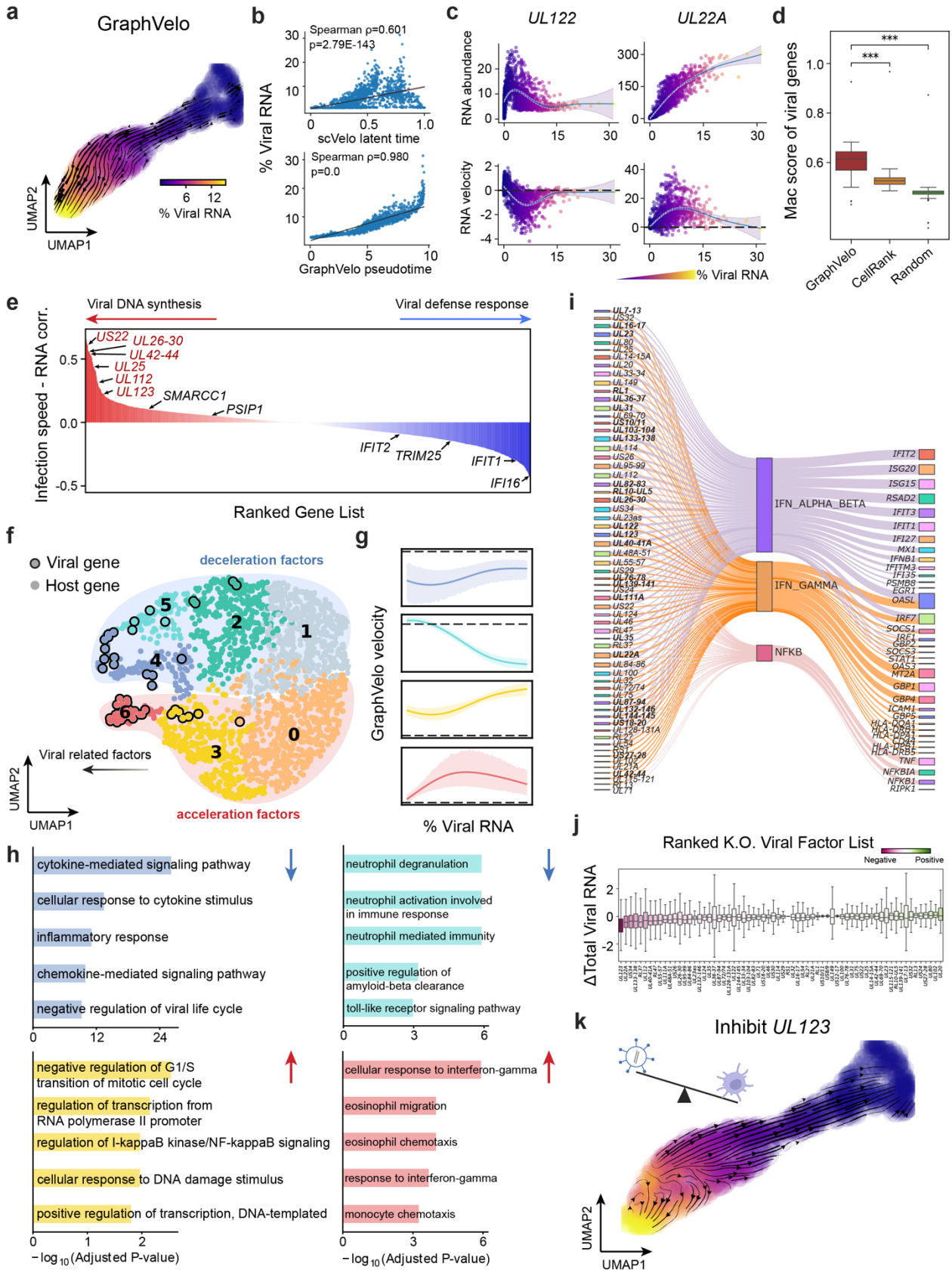
1069 **(i)** Activation of *Gata1* inhibitor TF *Spi1* lead to reversed velocity flows in gastrulation erythroid
1070 maturation investigated through in silico perturbation analyses on GraphVelo-based vector field.

1071 **(j)** Velocities derived from GraphVelo for the branching lineage in the hematopoiesis development
1072 and projected onto a pre-defined TSNE embedding. Directions of the projected cell velocities on
1073 TSNE are in agreement with the reported differentiation directions.

1074 **(k)** Terminal states identified by CellRank based on Markov chain formulation derived from
1075 GraphVelo velocities.

1076 **(l)** Phase portrait, velocity estimated by scVelo, refined velocity by GraphVelo, and gene
1077 expression of mature mRNA of identified rapid degradation genes (*NPR3, ANGPT1*). The cells
1078 were colored by the palantir pseudotime[31] in the phase portrait. The box plots showed cell-specific
1079 $\boldsymbol{\gamma}$ for cells divided into bins according to pseudotime ordering in the phase protrait.

1080


a
GraphVelo
UMAP2
UMAP1
6 12
% Viral RNA
b
Spearman ρ=0.601
p=2.79E-143
% Viral RNA
scVelo latent time
Spearman ρ=0.980
p=0.0
GraphVelo pseudotime
c
UL122
UL22A
RNA abundance
RNA velocity
% Viral RNA
d
Mac score of viral genes
***
***
GraphVelo
CellRank
Random
e
Viral DNA synthesis
Viral defense response
Infection speed - RNA corr.
US22
UL26-30
UL42-44
UL25
UL112
UL123
SMARCC1
PSIP1
IFIT2
TRIM25
IFIT1
IFI16
Ranked Gene List
f
Viral gene
Host gene
deceleration factors
Viral related factors
acceleration factors
UMAP2
UMAP1
g
GraphVelo velocity
% Viral RNA
i
IFN_ALPHA_BETA
IFN_GAMMA
NFKB
IFIT2
ISG20
ISG15
RSAD2
IFIT3
IFIT1
IFI27
MX1
IFNB1
IFITM3
IFI35
PSMB8
EGR1
OASL
IRF7
SOCS1
IRF1
GBP2
SOCS3
STAT1
OAS3
NT2A
GBP1
GBP4
ICAM1
GBP5
HLA-DQA1
HLA-DRB1
HLA-DPA1
HLA-DPB1
HLA-DRB5
TNF
NFKBIA
NFKBIZ
RIPK1
j
Ranked K.O. Viral Factor List
Negative
Positive
ΔTotal Viral RNA
k
Inhibit UL123
UMAP2
UMAP1
h
cytokine-mediated signaling pathway
cellular response to cytokine stimulus
inflammatory response
chemokine-mediated signaling pathway
negative regulation of viral life cycle
neutrophil degranulation
neutrophil activation involved in immune response
neutrophil mediated immunity
positive regulation of amyloid-beta clearance
toll-like receptor signaling pathway
negative regulation of G1/S transition of mitotic cell cycle
regulation of transcription from RNA polymerase II promoter
regulation of I-kappaB kinase/NF-kappaB signaling
cellular response to DNA damage stimulus
positive regulation of transcription, DNA-templated
cellular response to interferon-gamma
eosinophil migration
eosinophil chemotaxis
response to interferon-gamma
monocyte chemotaxis
−log10(Adjusted P-value)
−log10(Adjusted P-value)

1081

1082 **Figure 4. Using GraphVelo velocities to infer host-virus infection trajectory and identify**
1083 **host-pathogen interactions.**

1084 **(a)** Viral infection captured by the GraphVelo velocity field. Cells were colored by the percentage
1085 of viral RNA within a single cell.

1086 **(b)** Correlation between viral RNA percentage and pseudotime inferred by scVelo or GraphVelo.
1087 Spearman correlation coefficients and *P* values were shown.

1088 **(c)** Viral RNA velocities infered by GraphVelo along the viral RNA percentage axis. The black
1089 dot line highlights the zero velocity.

1090 **(d)** Boxplot summarizing the MacK scores of all viral genes calculated by GraphVelo, CellRank
1091 pseudotime kernel and random predictor.

1092 **(e)** Correlation between viral infection speed and RNA abundance. Genes were ranked by
1093 Spearman correlatioin coefficients. Host and viral genes that contribute to viral DNA synthesis
1094 were marked in the left side and those contribute to viral defense response were marked in the right
1095 side. Viral genes were highlighted in red.

1096 **(f)** UMAP representation of host and viral genes with distances defined by their dynamic
1097 expression patterns along the viral RNA percentage axis.

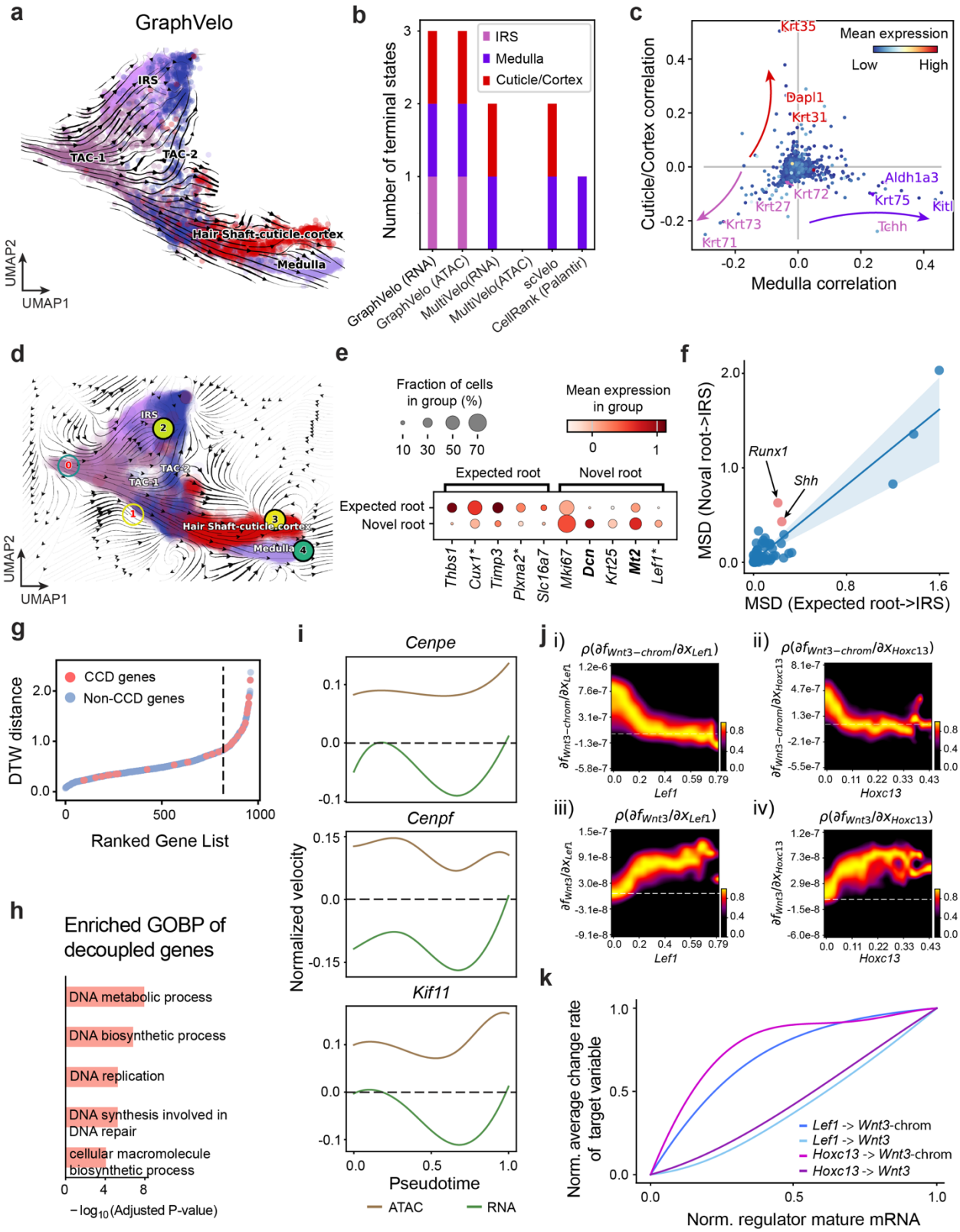
1098 **(g)** Example dynamic expression patterns within specific clusters (Leiden4, 5, 3, 6 from top to
1099 bottom) along the viral RNA percentage axis. Zero velocity was highlighted by black dot line.

1100 **(h)** GO enrichment of each cluster in (g).

1101 **(i)** Top host genes inhibited by each viral factor based on dynamo Jacobian analyses. Host effectors
1102 were organized by their involved pathways.

1103 **(j)** Dynamo prediction of total viral RNA change in response to in silico viral factor knockout.
1104 Viral factors were ranked by the mean of total viral RNA changes.

1105 **(k)** Vector field change resultant from infinitesmal inhibition of *UL123* during the viral infection
1106 process.


a
GraphVelo
IRS
TAC-1
TAC-2
Hair Shaft-cuticle.cortex
Medulla
UMAP2
UMAP1
b
Number of terminal states
IRS
Medulla
Cuticle/Cortex
GraphVelo (RNA)
GraphVelo (ATAC)
MultiVelo(RNA)
MultiVelo(ATAC)
scVelo
CellRank (Palantir)
c
Mean expression
Low
High
Cuticle/Cortex correlation
Medulla correlation
Krt35
Dapl1
Krt31
Aldh1a3
Krt72
Krt75
Krt27
Krt73
Krt71
Tchh
Kitl
d
IRS
TAC-2
TAC-1
Hair Shaft-cuticle.cortex
Medulla
UMAP2
UMAP1
e
Fraction of cells in group (%)
Mean expression in group
Expected root
Novel root
Thbs1
Cux1*
Timp3
Plxna2*
Slc16a7
Mki67
Dcn
Krt25
Mt2
Lef1*
f
MSD (Noval root->IRS)
MSD (Expected root->IRS)
Runx1
Shh
g
CCD genes
Non-CCD genes
DTW distance
Ranked Gene List
h
Enriched GOBP of decoupled genes
DNA metabolic process
DNA biosynthetic process
DNA replication
DNA synthesis involved in DNA repair
cellular macromolecule biosynthetic process
−log10(Adjusted P-value)
i
Cenpe
Cenpf
Kif11
Normalized velocity
Pseudotime
ATAC
RNA
j
i)
ii)
iii)
iv)
Lef1
Hoxc13
k
Norm. average change rate of target variable
Norm. regulator mature mRNA
Lef1 -> Wnt3-chrom
Lef1 -> Wnt3
Hoxc13 -> Wnt3-chrom
Hoxc13 -> Wnt3

1107

1108 **Figure 5. Inferring epigenome and transcriptome consistent dynamics in mouse hair follicle**
1109 **development using GraphVelo multi-omics velocities.**

1110 **(a)** GraphVelo velocity fields of mouse hair follicle development. Cells were colored by cell
1111 macrostates.

1112 **(b)** Number of terminal states predicted by CellRank using velocities inferred with different
1113 methods.

1114 **(c)** Driver genes along multiple lineages identified through CellRank.

1115 **(d)** Topological analyses of GraphVelo vector field identified novel root cells and attractors
1116 residing in three terminal states(IRS, hair shaft-cuticle cortex, and medulla).

1117 **(e)** Expression levels of marker genes in novel root cells and expected root cells. Markers identified
1118 by Ma et al.[61] were highlighted with stars and newly identified markers were highlighted in bold.

1119 **(f)** Regression results of MSD values along the transition path from the expected root or novel root
1120 to IRS. Two genes *Runx1* and *Shh* genes with large MSD originating from the novel root were
1121 highlighted.

1122 **(g)** DTW distance between RNA velocity and chromatin velocity of individual genes. CCD genes
1123 were colored in red. The dotted line indicates the elbow point separating the decoupled genes from
1124 the rest.

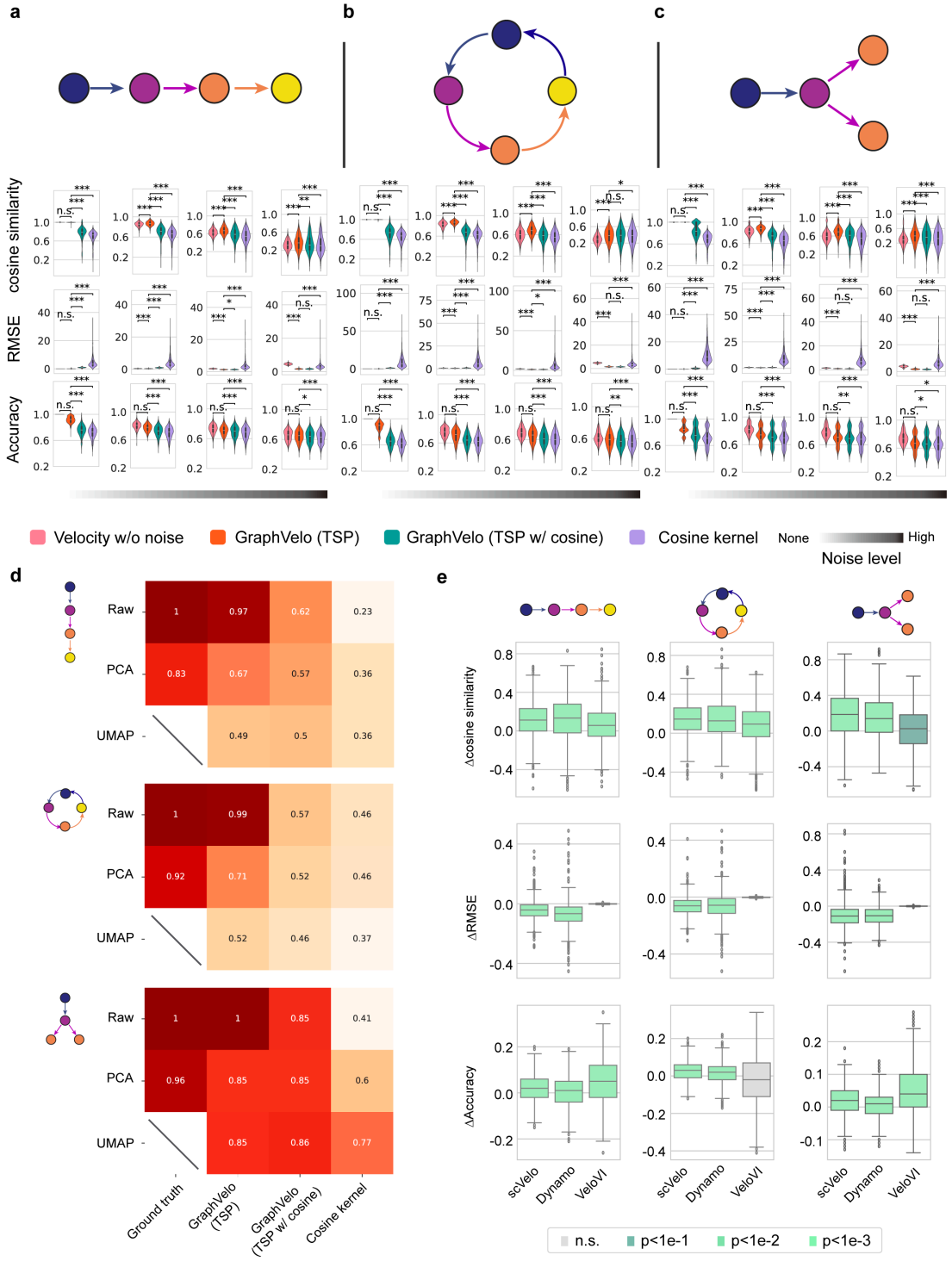
1125 **(h)** GO enrichment of decoupled genes in (g).

1126 **(i)** Line plot of nomarlized RNA and chromatin velocity along pseudotime for genes predicted by
1127 GraphVelo to have notable decoupling patterns. Chromatin velocity trends were colored as brown
1128 and RNA velocity trends were colored as green.

1129 **(j)** Heatmaps of Jacobian element distribution along the axis of regulator RNA abundance of four
1130 regulator effector circuits: i) *Lef1* versus *Wnt3* chromatin accessibilities. ii) *Hoxc13* versus *Wnt3*
1131 chromatin accessibilities. iii) *Lef1* versus *Wnt3* transcription. iv) *Hoxc13* versus *Wnt3*
1132 transcription.

1133 **(k)** Effective dose-response curves obtained from integrating the averaged Jacobian elements over
1134 the corresponding normalized regulator mature mRNA regulator level in (j).

Extended Data Figures:

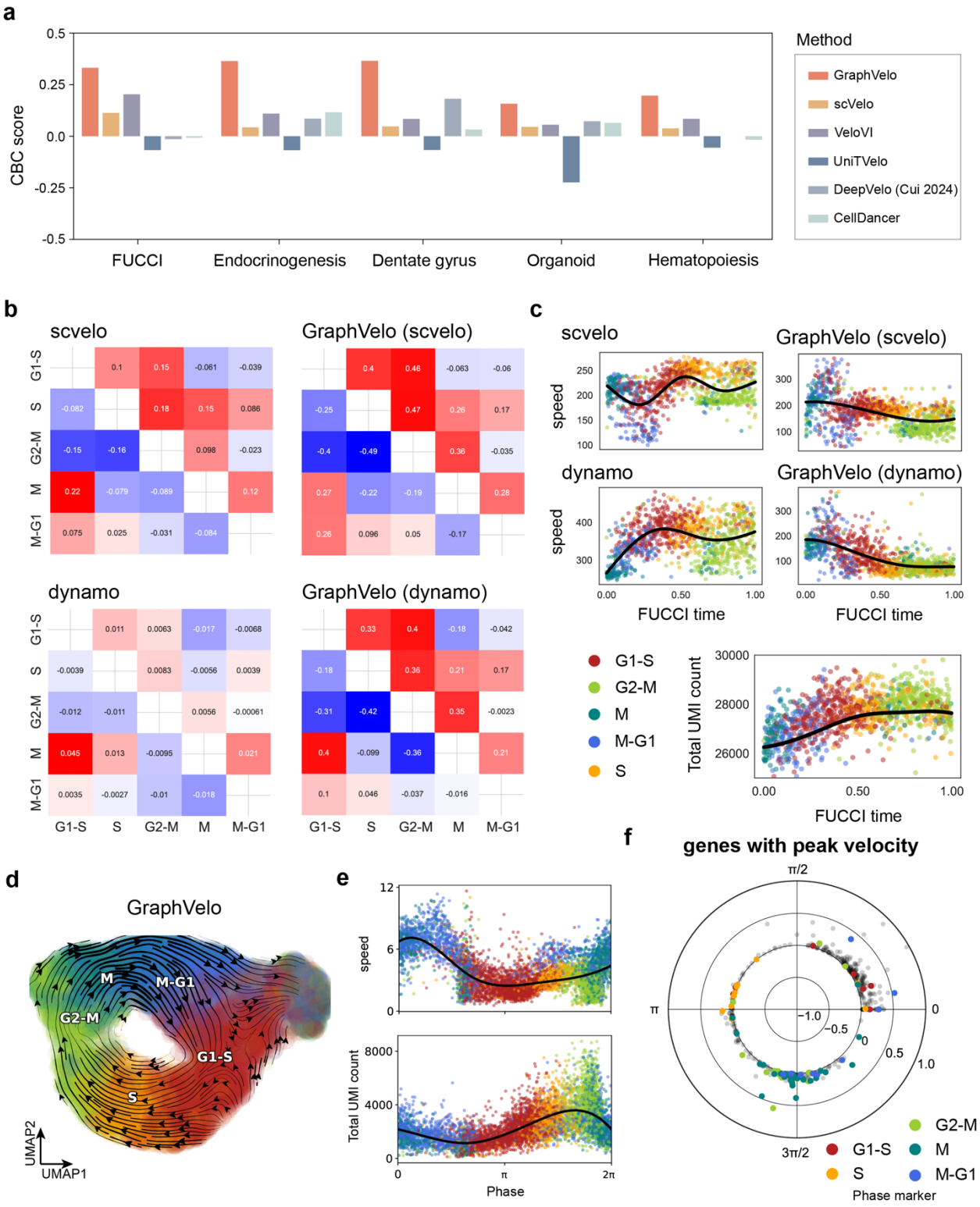


Extended Data Fig. 1. Refining RNA velocity from noisy simulation data or tradition splicing-based methods.

(a-c) Evaluation of on simulated scRNA-seq data under linear, cycling and bifurcating differentiation models with an increasing noise level in velocity vectors.

(d) Heatmap of the correlation of cell speed calculated as the norm of velocity vector with respect to the full-dimensional RNA velocity and the norm of velocity vectors projected to PCA or UMAP space using GraphVelo, GraphVelo with cosine regularization, cosine kernel.

(e) Boxplots of metric evaluations on GraphVelo correction to the original velocities estimated by scVelo, dynamo, and VeloVI. Box colors indicate the statistical test results for improvement relative to zero.



Extended Data Fig. 2. Quantitative benchmarks on GraphVelo across diverse biological systems.

(a) The average CBC scores across five benchmark datasets using different RNA velocity estimation methods.

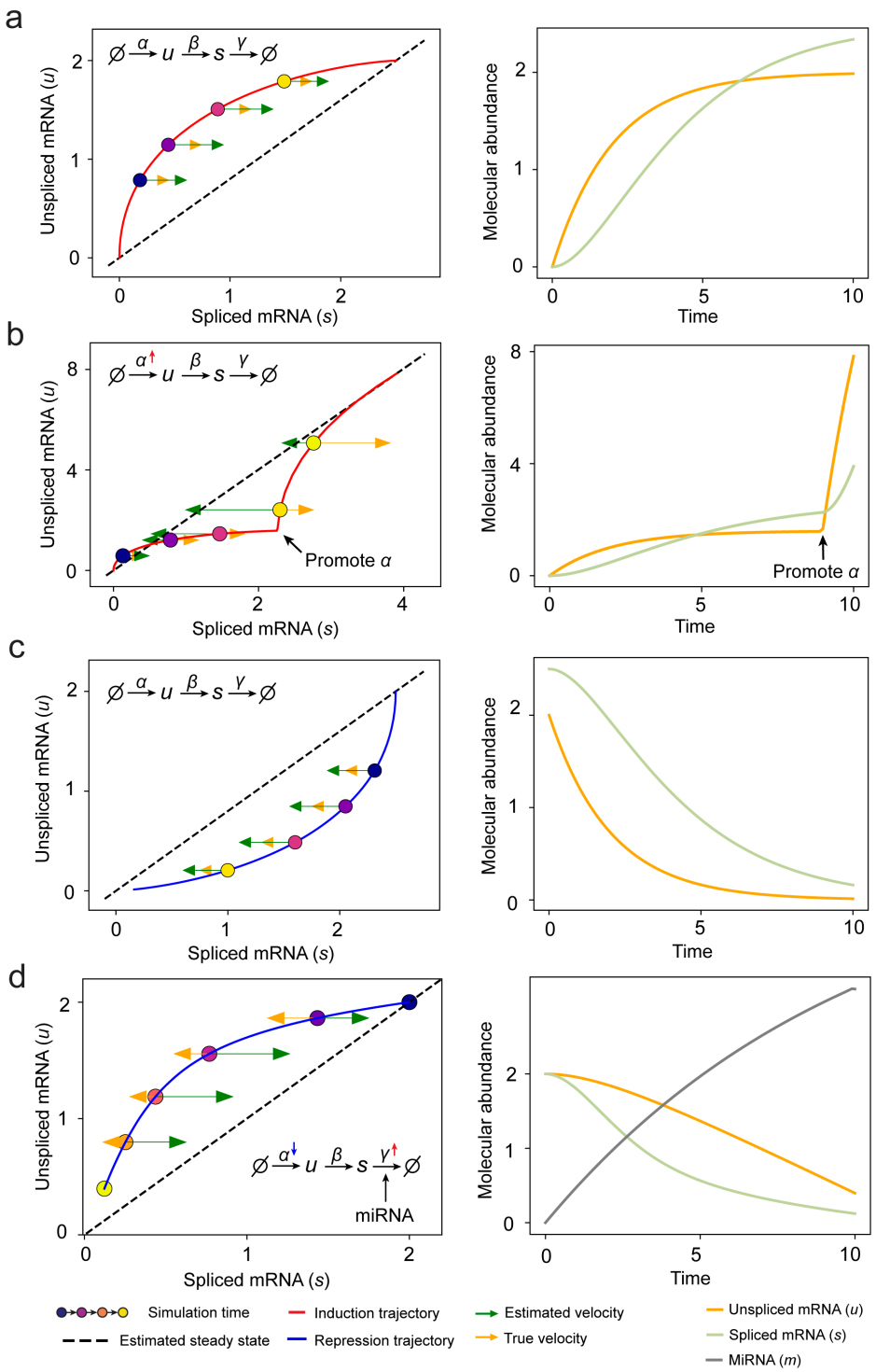
(b) Pairwise CBC scores between cell cycle phases (G1-S, S, G2-M, M, M-G1) using velocity vectors estimated by scVelo, dynamo, and their GraphVelo-refined outcomes.

(c) Cell cycle speed comparison along FUCCI time. GraphVelo-refined velocities identify the peak region within M-G1 phase and show a better match to total UMI distribution, reflecting more accurate scaling of cell cycle dynamics.

(d) Vector field reconstructed by GraphVelo on the A549 dataset, projected onto UMAP. Cells are colored by their cell cycle phases.

(e) Cell cycle speed (top) and total UMI count changes (bottom) plotted along the cell cycle phase. Peaks in velocity magnitude coincide with result in FUCCI data, particularly in M-G1.

(f) Polar plot of genes with peak RNA velocity, stratified by cell cycle phase, reflecting the cascade activation of phase marker genes.



Extended Data Fig. 3. RNA velocity inference on simulation results of splicing kinetics with constant or changed rates.

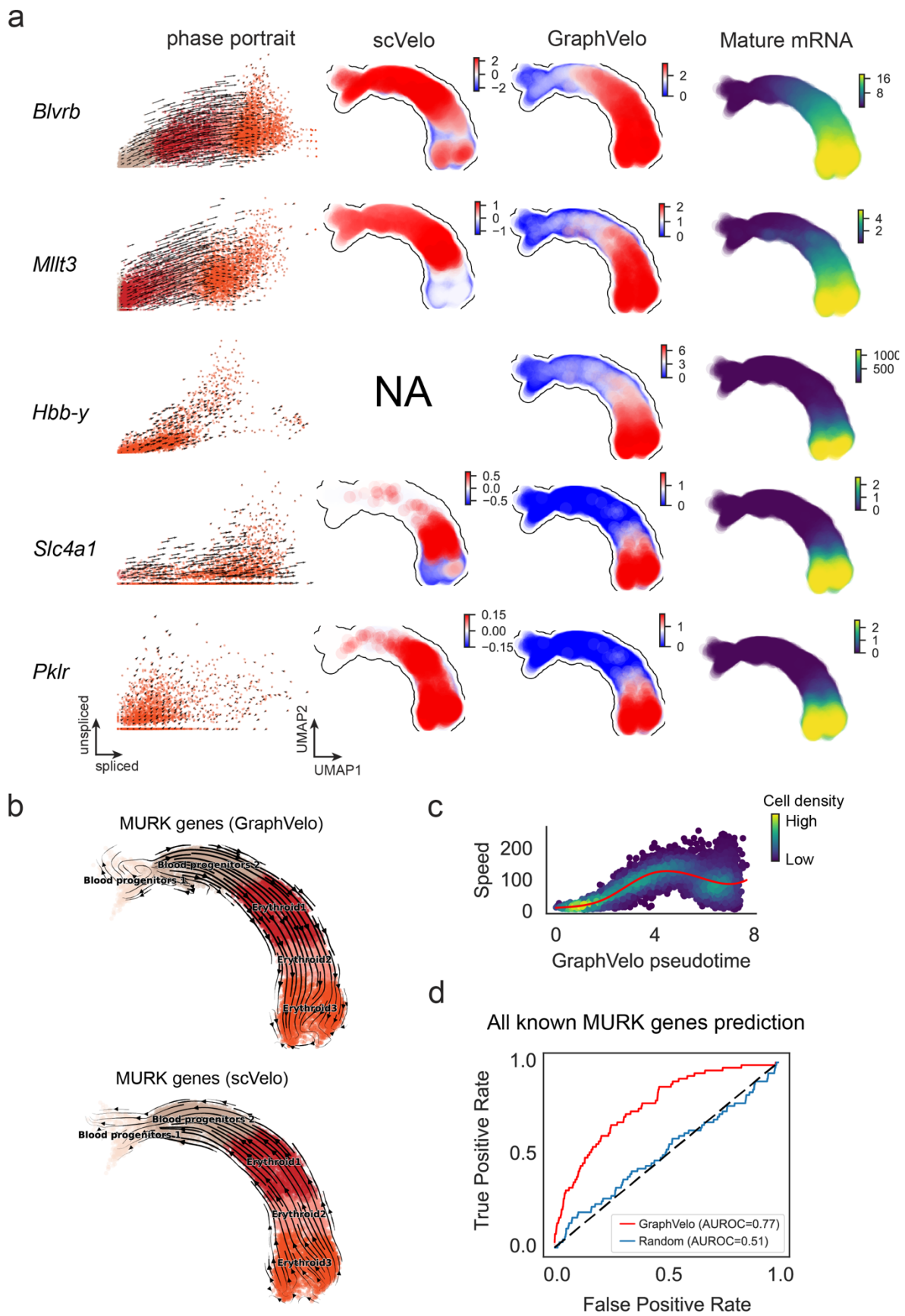
(a) Standard splicing kinetics along an induction trajectory in phase portrait and the reactions define how abundance levels of molecules change along simulation time. Right panel shows

corresponding trajectories over time (same below).

(b) Transcription burst along an induction trajectory in phase portrait. The transcription rate constant $\boldsymbol{\alpha}$ was promoted at specific time point.

(c) Standard splicing kinetics along a repression trajectory.

(d) Rapid time-varying degradation kinetics along repression trajectory. External signal promotes synthesis of microRNA, which enhances degradation of target mRNA resulting in a microRNA - dependent varying degradation rate "constant" $\boldsymbol{\gamma}$, and inhibits the target gene via a decreasing $\boldsymbol{\alpha}$.


a
phase portrait
scVelo
GraphVelo
Mature mRNA
Blvrb
Mllt3
Hbb-y
Slc4a1
Pklr
NA
unspliced
spliced
UMAP2
UMAP1
b
MURK genes (GraphVelo)
Blood progenitors 2
Blood progenitors 1
Erythroid1
Erythroid2
Erythroid3
MURK genes (scVelo)
c
Cell density
High
Low
Speed
GraphVelo pseudotime
d
All known MURK genes prediction
True Positive Rate
False Positive Rate
GraphVelo (AUROC=0.77)
Random (AUROC=0.51)

1180

1181 **Extended Data Fig. 4. GraphVelo extrapolation of RNA velocities of an extended list of gene**
1182 **set in mouse erythroid maturation dataset.**

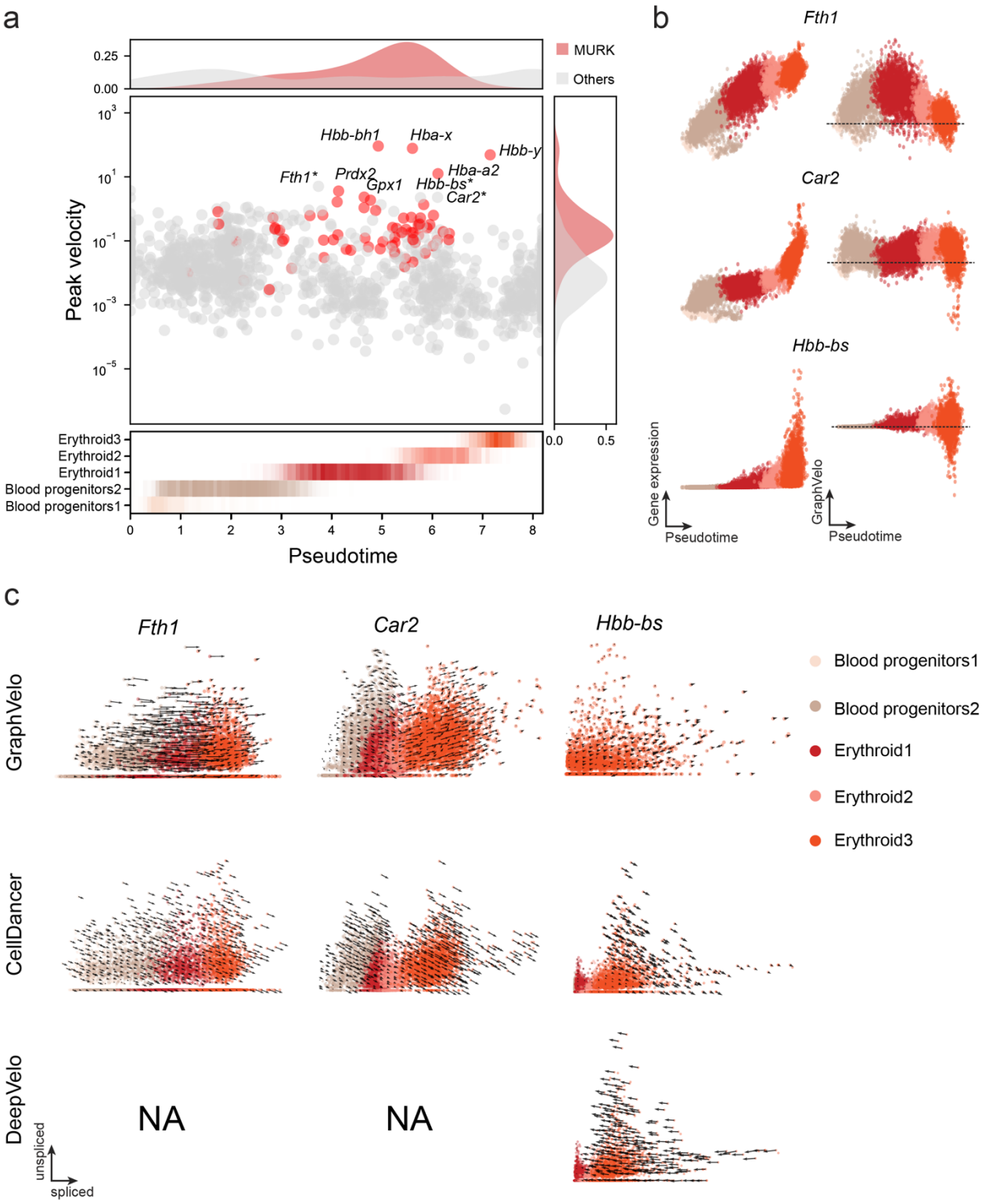
1183 **(a)** Phase portrait with GraphVelo vectors, velocity estimated by scVelo, refined velocity by
1184 GraphVelo, and gene expression of mature mRNA of a selected set of genes. Cells were colored
1185 by cell type, corresponding velocity, and mature mRNA abundance, respectively, and visualized
1186 on the phase portrait and UMAP, respectively.

1187 **(b)** Velocities of MURK genes derived from GraphVelo and scVelo for gastrulation erythroid
1188 maturation cells projected to a predefined UMAP representation. Note that RNA velocity
1189 estimated by scVelo was in a reverse flow, possibly influenced by transcription burst events.

1190 **(c)** Cell speed distribution along the vector field pseudotime axis. Cells were colored by local
1191 density and red line indicates the fitted curve.

1192 **(d)** Receiver operating curve analyses of MURK gene prediction based on dynamo accelerations
1193 analysis using GraphVelo as input, in contrast to a random predictor. AUC, area under curve.

1194

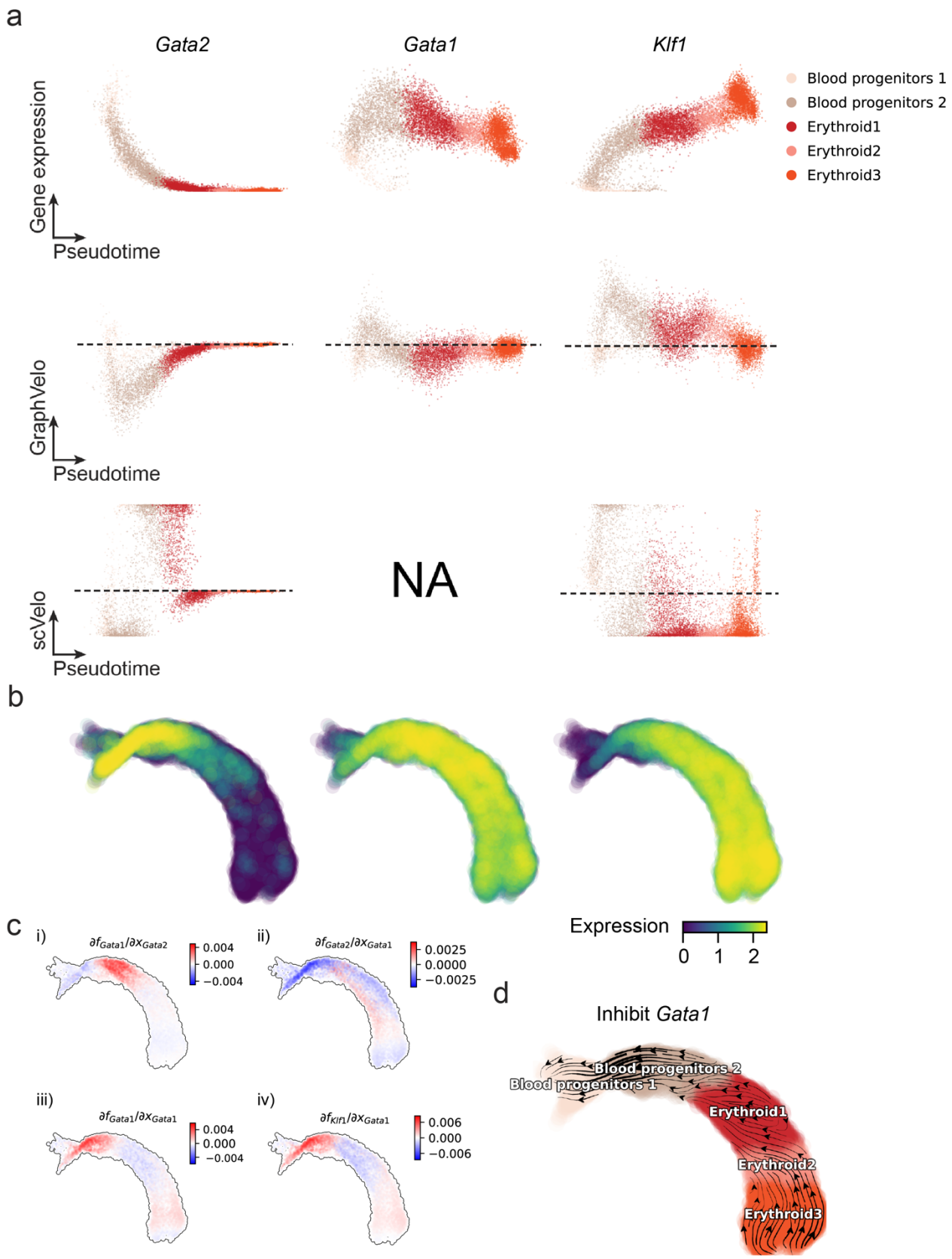


Extended Data Fig. 5. GraphVelo reveals potential MURK genes with larger peak velocity magnitude during erythroid maturation.

(a) Peak RNA velocity across pseudotime for highly variable genes. MURK genes (red) show significantly higher peak velocity magnitudes than other genes (gray), particularly in late-stage erythroid cells. Top peak velocity genes are annotated and *Fth1*, *Car2*, *Hbb-bs* are suspected as potential MURK genes and highlighted by star*. Density plots above and to the side illustrate the distributions of pseudotime and peak velocity, respectively. The heatmap below shows cell density along pseudotime by cell states.

(b) Expression and corresponding velocities of *Fth1*, *Car2*, *Hbb-bs* along the vector field-based pseudotime in mouse erythroid data.

(c) Phase portraits comparing velocity vectors estimated by GraphVelo, CellDancer, and DeepVelo for selected potential MURK genes. GraphVelo reveals well-directed and temporally aligned transcriptional dynamics, while alternative methods (particularly DeepVelo) fail to provide interpretable velocity fields for key genes with altered kinetics.


a
Gata2
Gata1
Klf1
Blood progenitors 1
Blood progenitors 2
Erythroid1
Erythroid2
Erythroid3
Gene expression
Pseudotime
GraphVelo
Pseudotime
scVelo
Pseudotime
NA
b
Expression
0
1
2
c
i)
∂f_Gata1/∂x_Gata2
0.004
0.000
−0.004
ii)
∂f_Gata2/∂x_Gata1
0.0025
0.0000
−0.0025
iii)
∂f_Gata1/∂x_Gata1
0.004
0.000
−0.004
iv)
∂f_Klf1/∂x_Gata1
0.006
0.000
−0.006
d
Inhibit Gata1
Blood progenitors 2
Blood progenitors 1
Erythroid1
Erythroid2
Erythroid3

1213

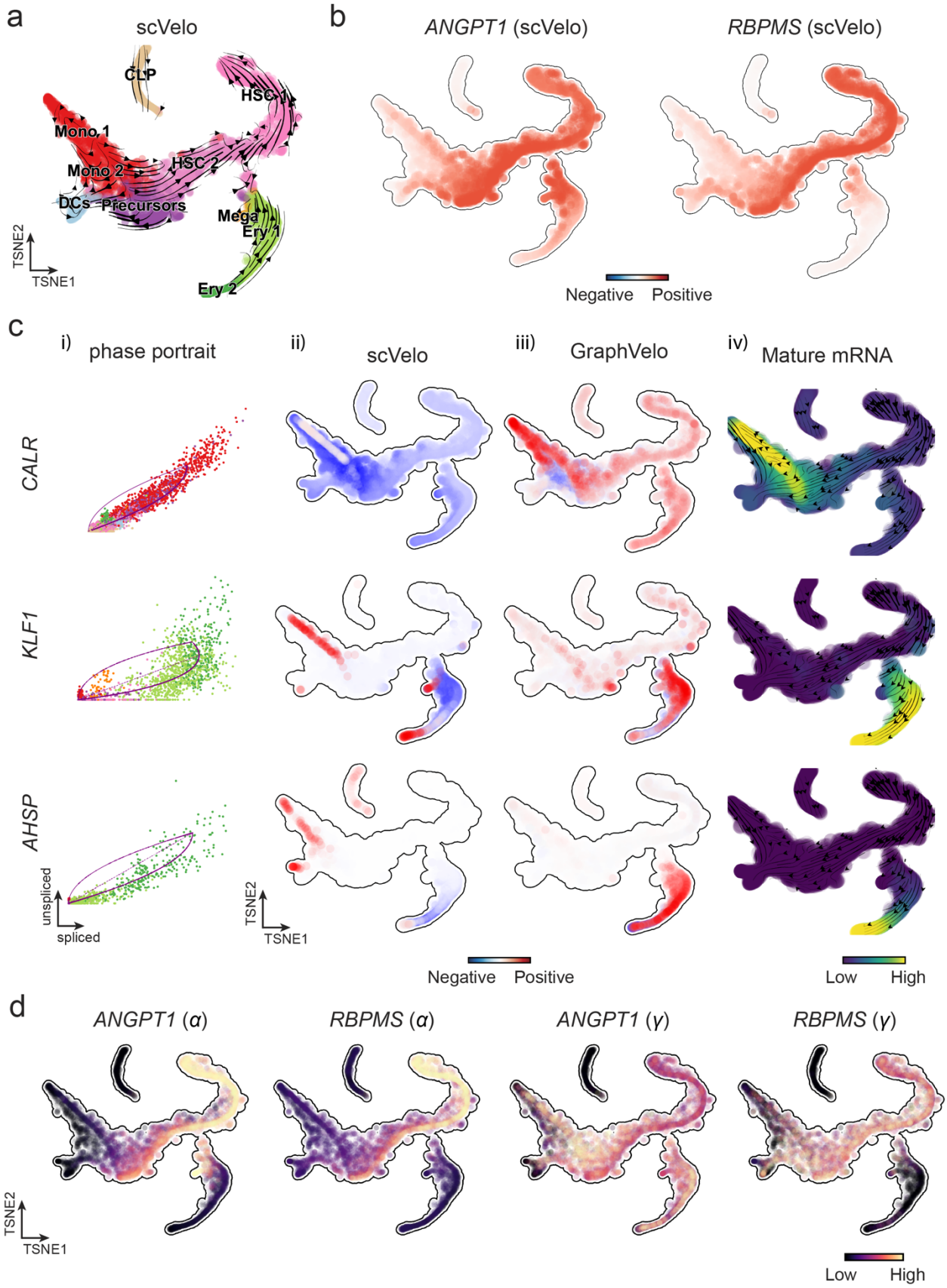
1214 **Extended Data Fig. 6. Regulatory cascades of driver TFs in mouse erythroid data.**

1215 **(a)** Expression and corresponding velocities of *Gata2*, *Gata1*, *Klf1* along the vector field-based
1216 pseudotime in mouse erythroid data.

1217 **(b)** Gene expression of *Gata2*, *Gata1*, *Klf1* on the UMAP space.

1218 **(c)** Molecular mechanisms of driver TFs underlying erythroid lineage commitment based on
1219 Jacobian analyses. (i) *Gata2* activates *Gata1*. (ii) Repression of *Gata2* by *Gata1*. (iii) Self-
1220 activation of *Gata1*. (iv) *Gata1* activates *Klf1*.

1221 **(d)** In silico perturbation analyses on GraphVelo-based vector field to examine the role of *Gata1*
1222 in gastrulation erythroid maturation.


a
scVelo
CLP
HSC 1
Mono 1
Mono 2
HSC 2
DCs
Precursors
Mega
Ery 1
Ery 2
TSNE2
TSNE1
b
ANGPT1 (scVelo)
RBPMS (scVelo)
Negative
Positive
c
i) phase portrait
ii) scVelo
iii) GraphVelo
iv) Mature mRNA
CALR
KLF1
AHSP
unspliced
spliced
TSNE2
TSNE1
Negative
Positive
Low
High
d
ANGPT1 (α)
RBPMS (α)
ANGPT1 (γ)
RBPMS (γ)
TSNE2
TSNE1
Low
High

1223

1224 **Extended Data Fig. 7. RNA velocity estimated by scVelo and refined by GraphVelo from a**
1225 **hematopoiesis dataset.**

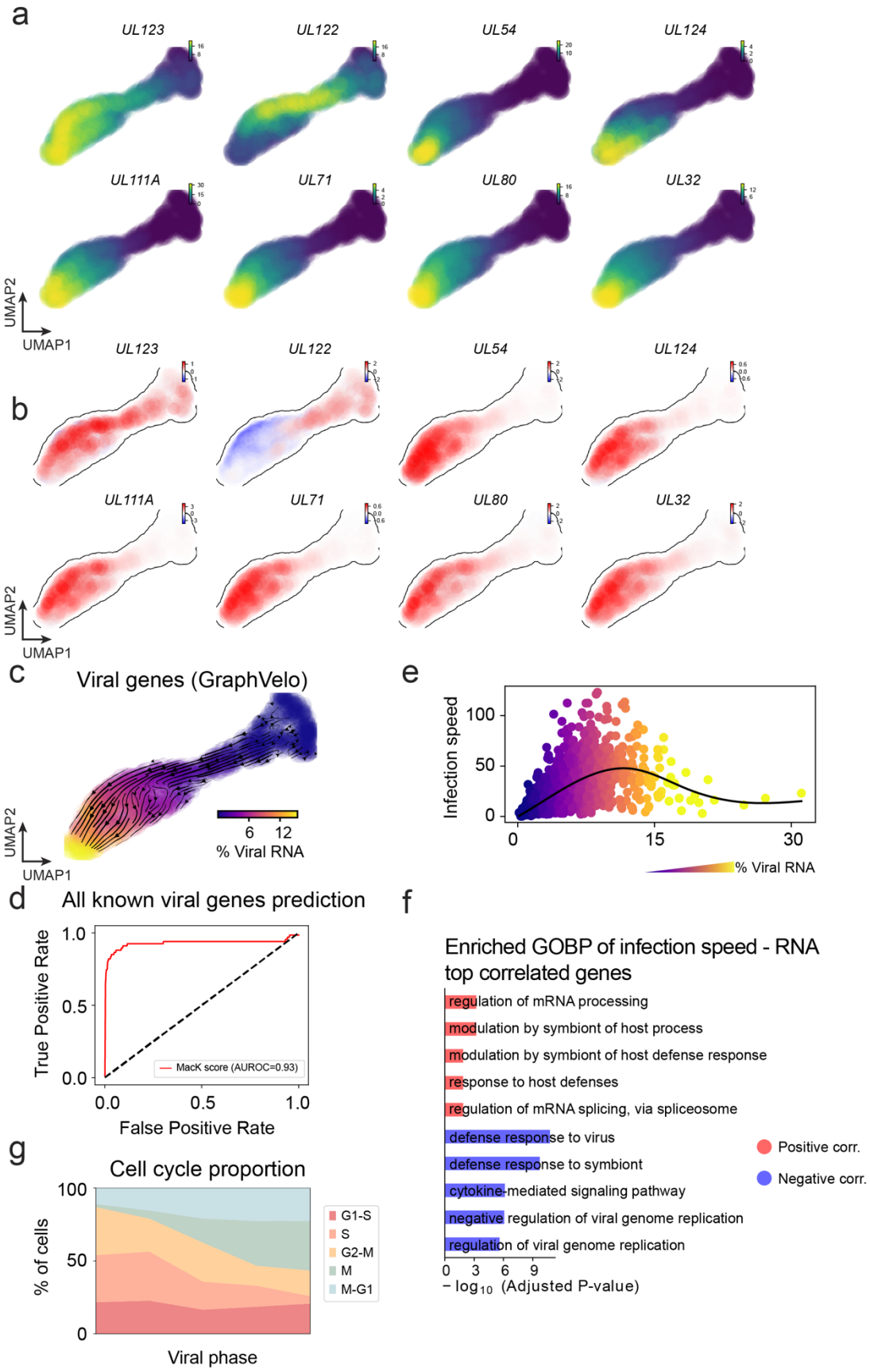
1226 **(a)** Velocities derived from scVelo in the hematopoiesis development and projected to a pre-
1227 defined TSNE embedding.

1228 **(b)** TSNE visualization of the RNA velocity estimated by scVelo of *ANGPT1* and *RBPMS* genes.

1229 **(c)** Scatter plots of: i) phase portrait, ii) velocities estimated by scVelo, iii) refined velocities by
1230 GraphVelo, and iv) mature mRNA expression of transcription burst genes (e.g. *CALR*, *KLF1* and
1231 *AHSP*). Cells were colored by cell type, corresponding velocity, and mature mRNA abundance,
1232 respectively, and visualized on the phase portrait and TSNE, respectively.

1233 **(d)** Cell-specific transcription rate constant $\boldsymbol{\alpha}$ and degradation rate constant $\boldsymbol{\gamma}$ of rapid degradation
1234 genes *ANGPT1* and *RBPMS* visualized on the TSNE space, which is consistent with simulation
1235 result in Extended Data Fig. 3d.

1236


a
UL123
UL122
UL54
UL124
UL111A
UL71
UL80
UL32
UMAP2
UMAP1
b
UL123
UL122
UL54
UL124
UL111A
UL71
UL80
UL32
UMAP2
UMAP1
c
Viral genes (GraphVelo)
UMAP2
UMAP1
6
12
% Viral RNA
e
Infection speed
100
50
0
0
15
30
% Viral RNA
d
All known viral genes prediction
True Positive Rate
1.0
0.5
0.0
0.0
0.5
1.0
MacK score (AUROC=0.93)
False Positive Rate
f
Enriched GOBP of infection speed - RNA top correlated genes
regulation of mRNA processing
modulation by symbiont of host process
modulation by symbiont of host defense response
response to host defenses
regulation of mRNA splicing, via spliceosome
defense response to virus
defense response to symbiont
cytokine-mediated signaling pathway
negative regulation of viral genome replication
regulation of viral genome replication
0
3
6
9
− log10 (Adjusted P-value)
Positive corr.
Negative corr.
g
Cell cycle proportion
% of cells
100
50
0
G1-S
S
G2-M
M
M-G1
Viral phase

1237

1238 **Extended Data Fig. 8. Dynamic patterns of gene trajectories during HCMV infection.**

1239 **(a)** Gene expression of selected viral factors on the UMAP space.

1240 **(b)** Corresponding velocities inferred by GraphVelo for viral factors in (a).

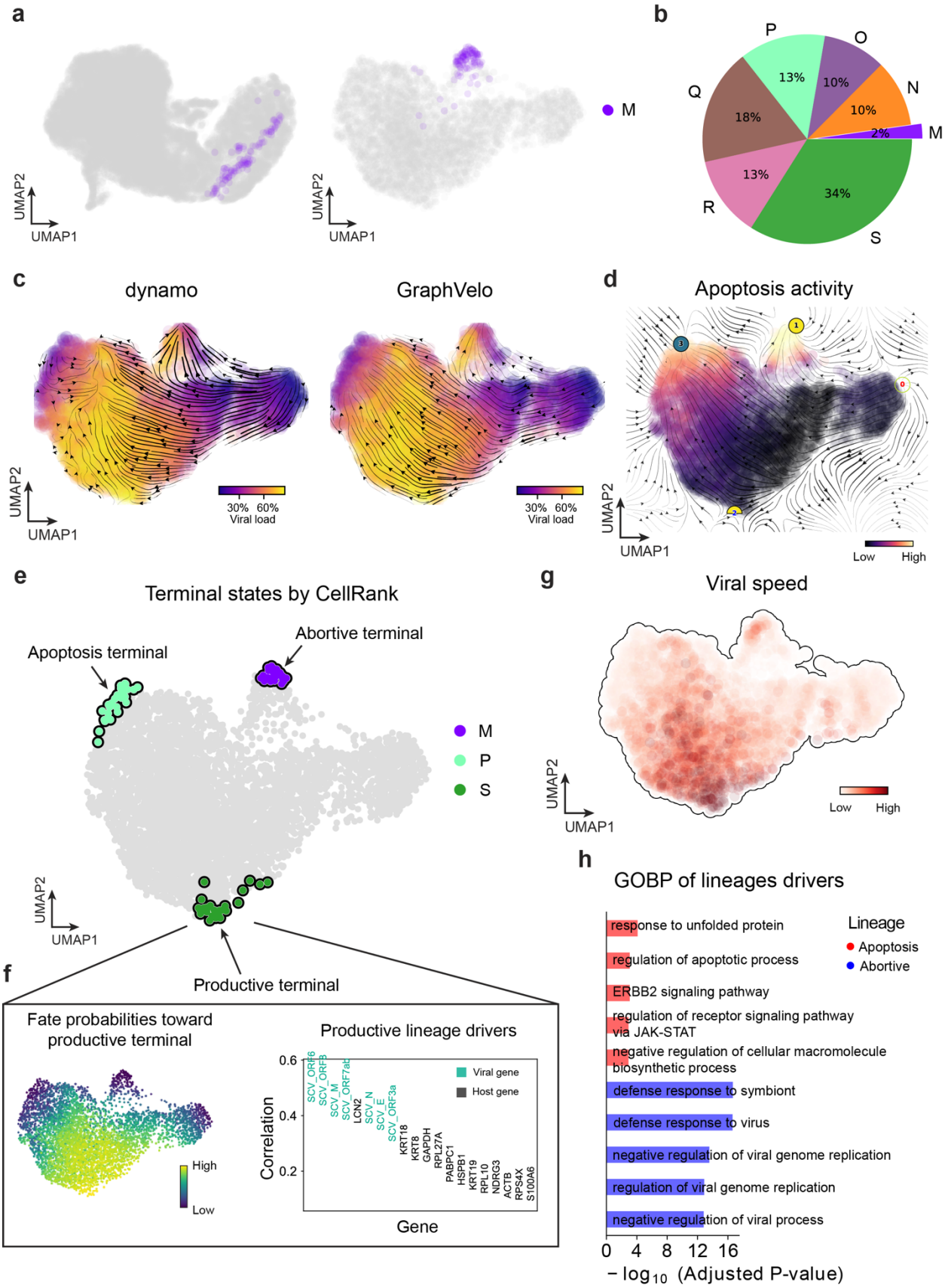
1241 **(c)** Viral RNA velocities derived from GraphVelo for infected cells, projected to the UMAP
1242 representation only using viral genes.

1243 **(d)** Receiver operating curve (ROC) analyses of MacK scores when using all detected viral genes
1244 as the gold standard.

1245 **(e)** Viral infection speed versus viral load. Cells were colored by the percentage of viral RNA.

1246 **(f)** GO enrichment analyses of top host genes correlated with infection speed.

1247 **(g)** Shifts in cell proportions across different cell cycle phases along the viral phase trajectory.
1248 Cells were divided into bins according to viral RNA percentage. The result indicates a decreasing
1249 S phase population corresponding to infection-driven transcriptional alterations of the cell cycle.
1250


a
UMAP2
UMAP1
M
b
P
O
Q
N
M
R
S
13%
10%
10%
2%
18%
13%
34%
c
dynamo
GraphVelo
30% 60%
Viral load
d
Apoptosis activity
Low
High
e
Terminal states by CellRank
Apoptosis terminal
Abortive terminal
Productive terminal
M
P
S
f
Fate probabilities toward productive terminal
High
Low
Productive lineage drivers
Correlation
Gene
Viral gene
Host gene
SCV_ORF6
SCV_ORF8
SCV_M
SCV_ORF7ab
LCN2
SCV_N
SCV_E
SCV_ORF3a
KRT18
KRT8
GAPDH
RPL27A
PABPC1
HSPB1
KRT19
RPL10
NDRG3
ACTB
RPS4X
S100A6
g
Viral speed
Low
High
h
GOBP of lineages drivers
Lineage
Apoptosis
Abortive
response to unfolded protein
regulation of apoptotic process
ERBB2 signaling pathway
regulation of receptor signaling pathway via JAK-STAT
negative regulation of cellular macromolecule biosynthetic process
defense response to symbiont
defense response to virus
negative regulation of viral genome replication
regulation of viral genome replication
negative regulation of viral process
0 4 8 12 16
− log10 (Adjusted P-value)

1251

1252 **Extended Data Fig. 9. GraphVelo infers multiple terminals driven by viral-host interactions.**

1253 **(a)** UMAP coordinates computed from all cell populations in the original study[55] (left) and from
1254 the infected subpopulation alone (right). Cells belonging to cluster M are highlighted in purple.

1255 **(b)** Proportion of cluster M within all infected cell population.

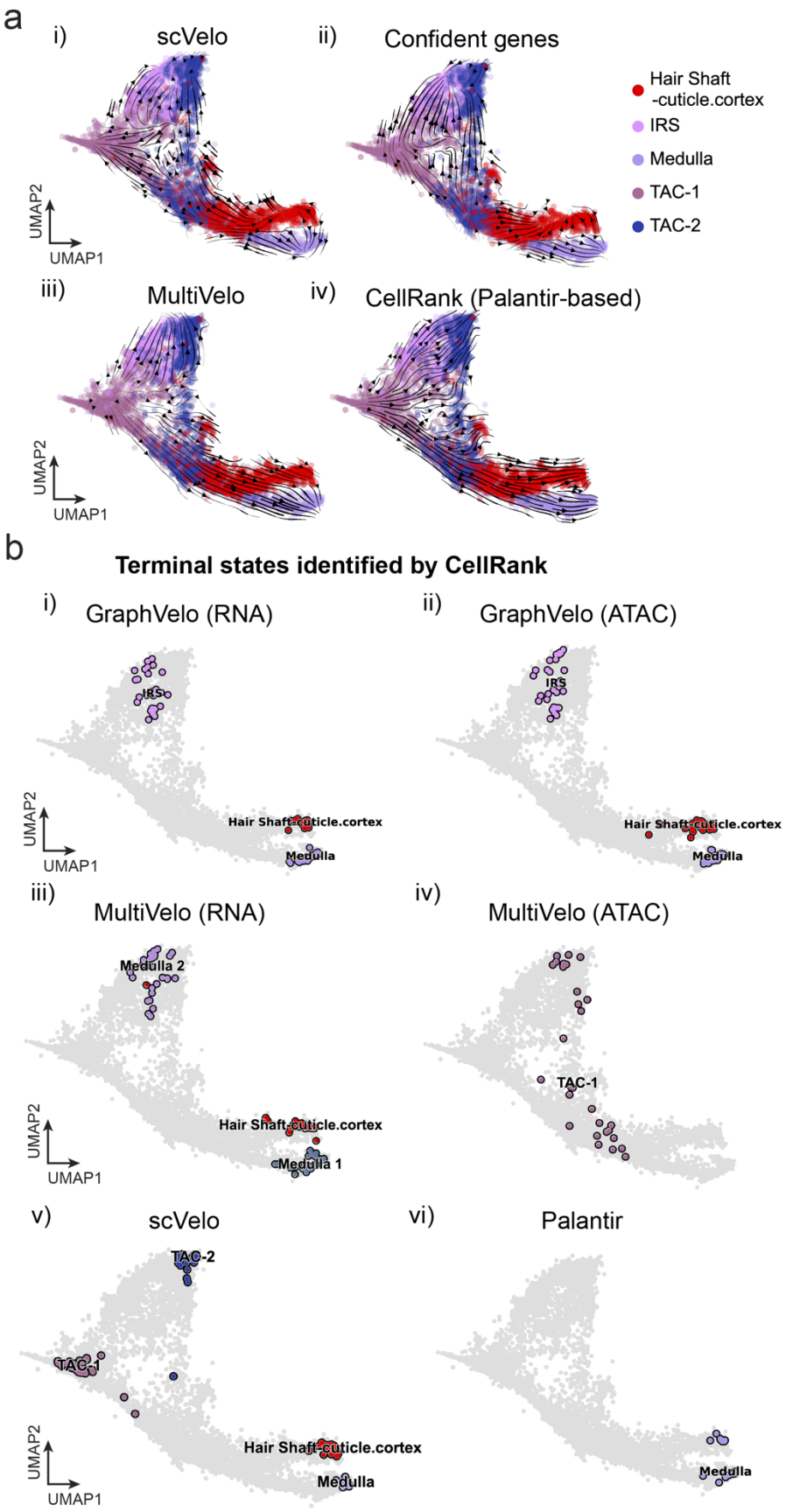
1256 **(c)** Velocities derived from dynamo (left) and GraphVelo (right) in the infected cells and projected
1257 to UMAP embedding. Cells are colored by viral load.

1258 **(d)** Topological analyses of GraphVelo vector field identified initial position with low viral RNA
1259 load, saddle point with high viral load (as shown in panel c) and attractors residing in terminal
1260 states with high apoptosis activity.

1261 **(e)** Terminal states identified by CellRank using GraphVelo-corrected RNA velocity.

1262 **(f)** Fate probabilities toward the productive terminal state (left) and top 20 lineage-correlated genes
1263 identified (right) based on GraphVelo-based CellRank analysis.

1264 **(g)** GO enrichment analyses of top lineage-driver genes along two distinct cell death trajectories.
1265


a
i)
scVelo
ii)
Confident genes
Hair Shaft
-cuticle.cortex
IRS
Medulla
TAC-1
TAC-2
UMAP2
UMAP1
iii)
MultiVelo
iv)
CellRank (Palantir-based)
UMAP2
UMAP1
b
Terminal states identified by CellRank
i)
GraphVelo (RNA)
IRS
Hair Shaft-cuticle.cortex
Medulla
UMAP2
UMAP1
ii)
GraphVelo (ATAC)
IRS
Hair Shaft-cuticle.cortex
Medulla
iii)
MultiVelo (RNA)
Medulla 2
Hair Shaft-cuticle.cortex
Medulla 1
UMAP2
UMAP1
iv)
MultiVelo (ATAC)
TAC-1
v)
scVelo
TAC-2
TAC-1
Hair Shaft-cuticle.cortex
Medulla
UMAP2
UMAP1
vi)
Palantir
Medulla

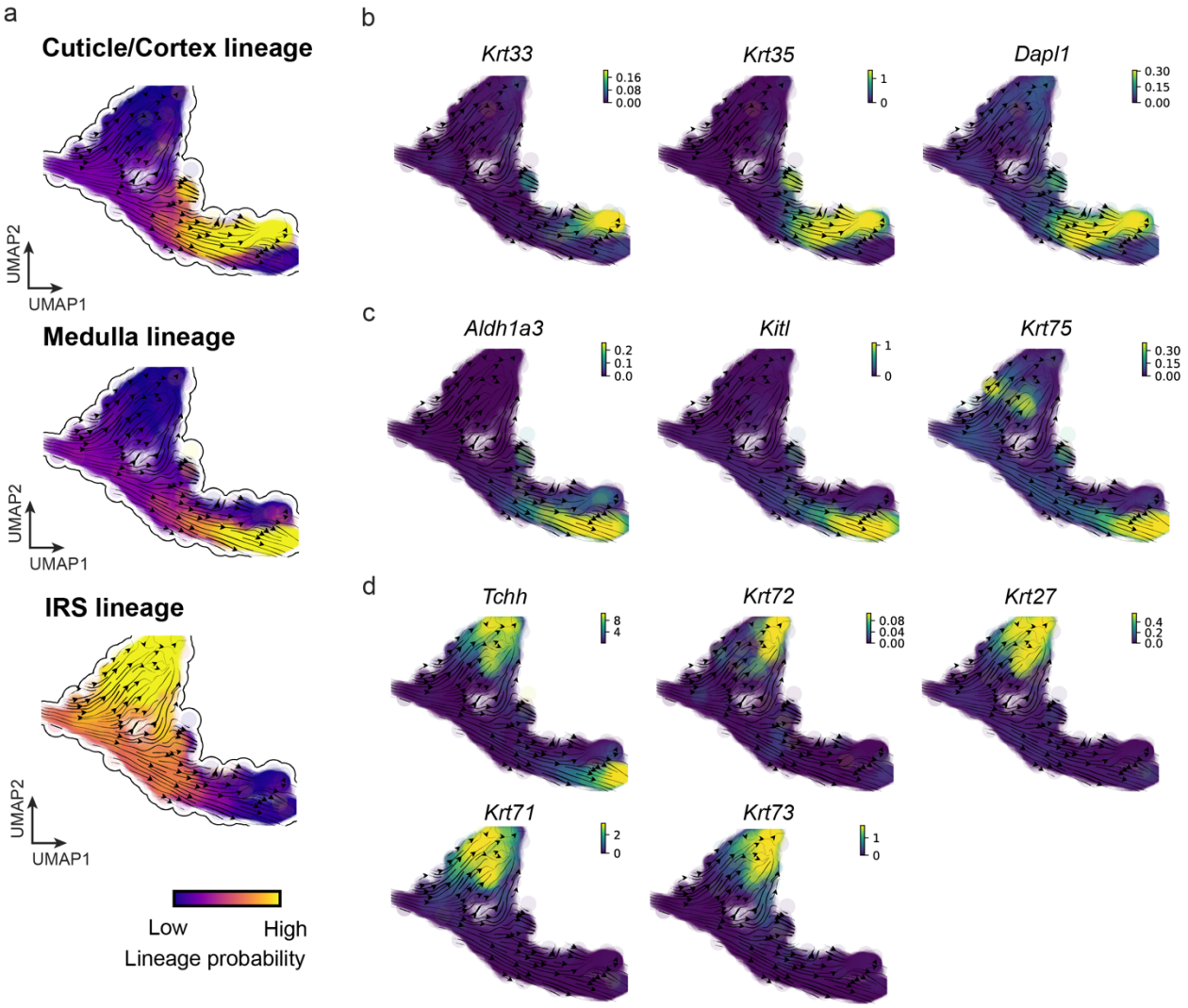
1266

1267 **Extended Data Fig. 10. GraphVelo estimation of modality-consistent vector field based on**
1268 **multi-omics velocities.**

1269 **(a)** Velocity field projected to the pre-defined UMAP representation with different methods or
1270 gene sets: i) scVelo RNA velocity based on all velocity genes; ii) scVelo RNA velocity based on
1271 confident velocity genes filterred by dynamo criteria; iii) MultiVelo RNA velocity based on all
1272 velocity genes; iv) CellRank pseudotime kernel based on Palantir pseudotime.

1273 **(b)** Terminal states identified by CellRank using different representation and corresponding
1274 velocity: i) GraphVelo-corrected RNA velocity; ii) GraphVelo-computed chromatin velocity; iii)
1275 MultiVelo-computed RNA velocity; iv) MultiVelo-computed chromatin velocity; v) scVelo-
1276 computed RNA velocity; vi) Palantir-based pseudotime kernel.

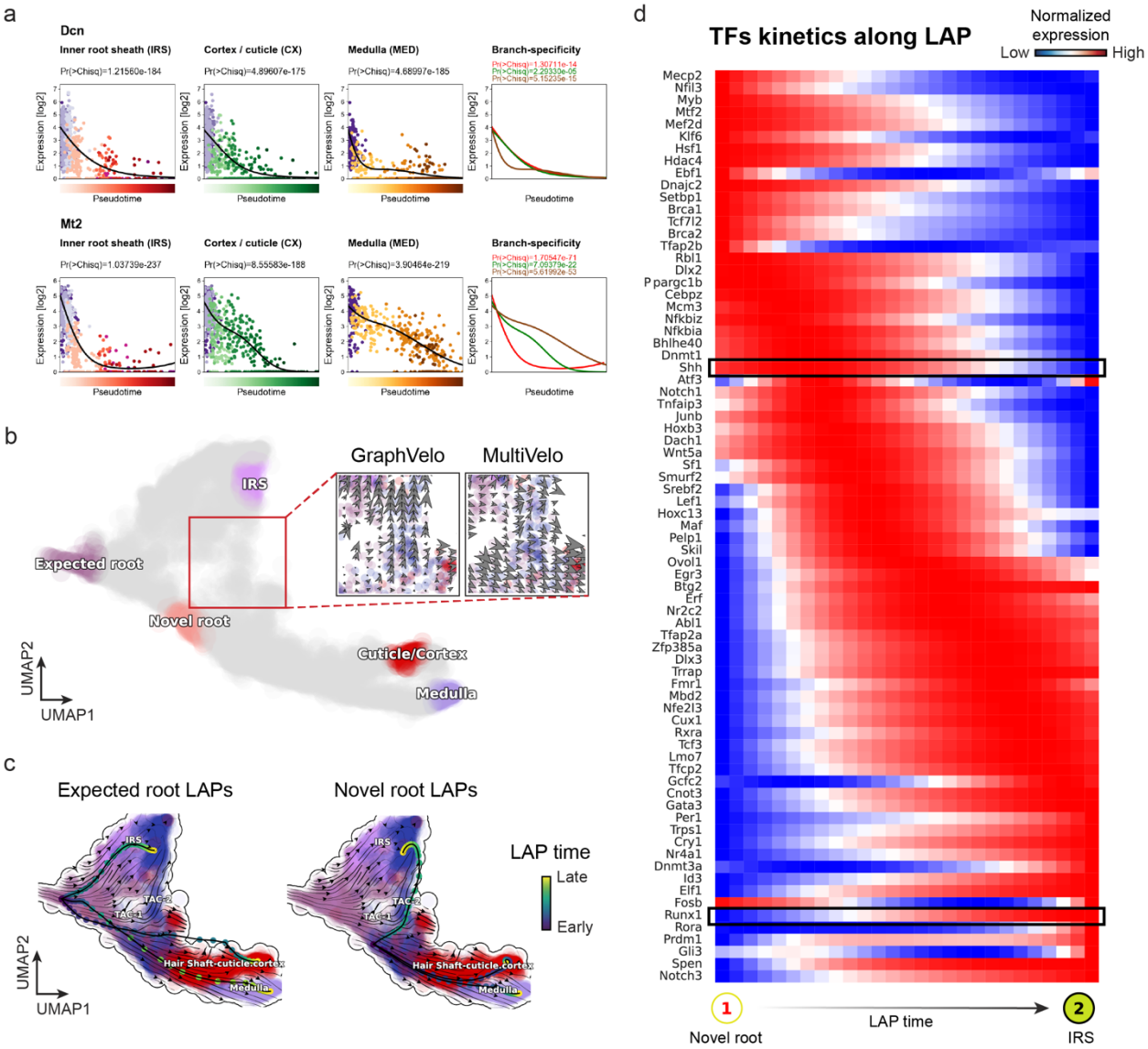
1277



Extended Data Fig. 11. Reconstructed lineage commitment during mouse hair follicle differentiation using GraphVelo velocities.

(a) Lineage commitment probablity of each terminal cell type on the UMAP vector field.

(b-d) Gene expression distribution of markers for cuticle/cortex, medulla and IRS lineages on the UMAP vector field.



Extended Data Fig. 12. Novel root identified with multi-omic vector field from GraphVelo velocities.

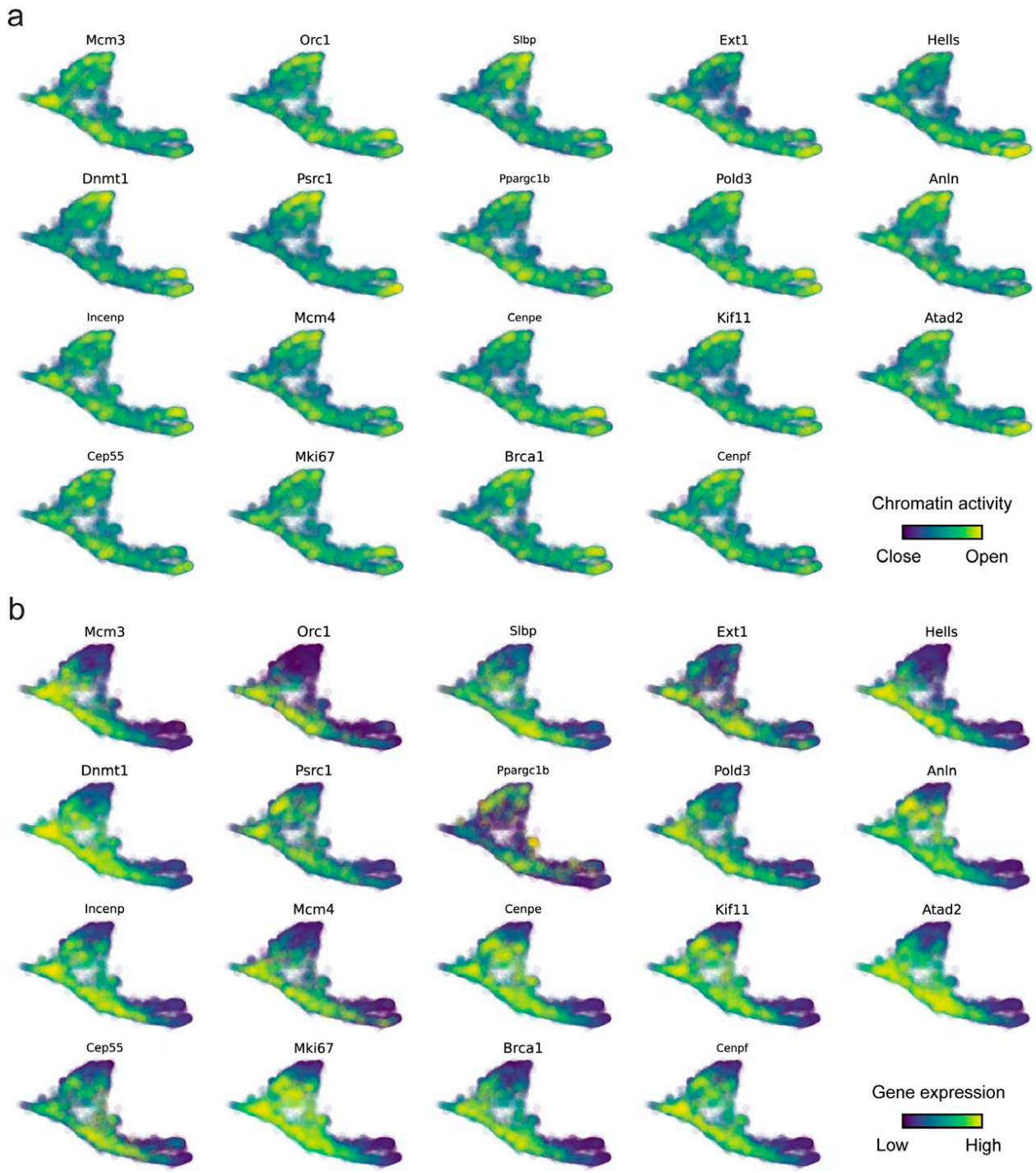
(a) *Dcn*, *Mt2* expression dynamics during anagen hair follicle keratinocytes[62].

(b) Regions identified by topological analysis. Insert are ummarized cell-state transition vectors calculated by GraphVelo and MultiVelo along the path from the novel root to IRS and projected onto the UMAP representation.

(c) Predicted developmental LAPs from expected root or novel root to to each of the terminal cell types in the UMAP embedding. Color of the node along the paths indicates the LAP transition time.

1295 **(d)** TF expression profiles along the LAP from novel root to IRS. *Shh* decays, alongside the
1296 induction of the *Runx1* gene.

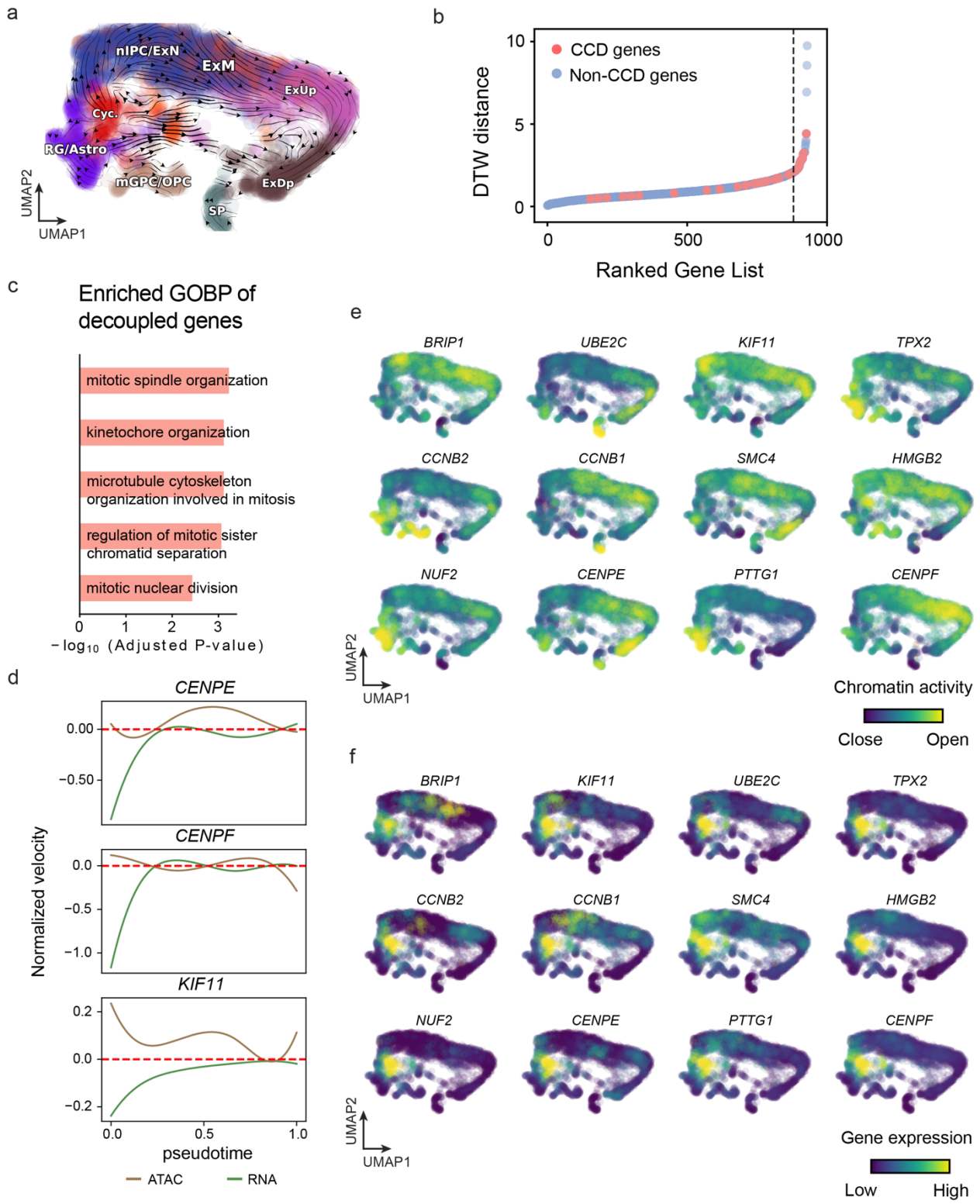
1297



Extended Data Fig. 13. Genomic patterns of decoupled CCD genes.

(a) Chromatin activity of decoupled CCD genes.

(b) Gene expression of decoupled CCD genes.



Extended Data Fig. 14. GraphVelo inferrence on epigenome and transcriptome decoupling dynamics in fetal human brain.

1305 **(a)** GraphVelo velocity field colored by cell macrostates.

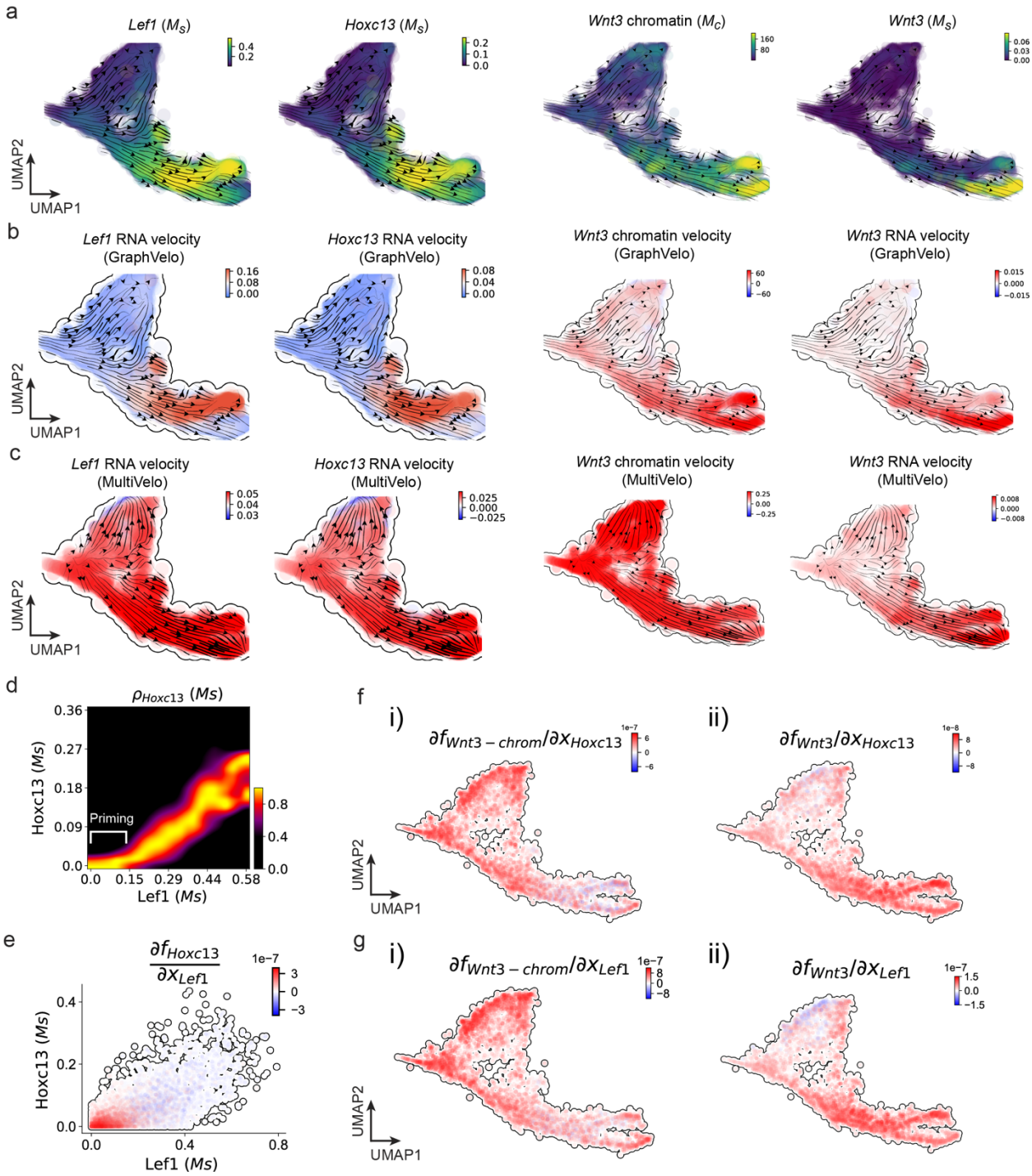
1306 **(b)** DTW distance between RNA velocity and chromatin velocity of individual genes as a measure
1307 of the coupling/decoupling status. CCD genes were colored in red. The dotted line indicates the
1308 elbow point, with the decoupled genes on its right.

1309 **(c)** GO enrichment of decoupled genes in (c).

1310 **(d)** Line plot of nomarlized RNA and chromatin velocity along dynamo vector field pseudotime
1311 for genes predicted by GraphVelo to have notable decoupling patterns. Chromatin velocity trends
1312 were colored in brown and RNA velocity trends were colored in green.

1313 **(e)** Chromatin activity of decoupled CCD genes.

1314 **(f)** Gene expression of decoupled CCD genes.



Extended Data Fig. 15. GraphVelo prediction on the regulatory mechanisms of *Lef1-Hoxc13-Wnt3* circuit during mouse skin development.

(a) Distributions of *Lef1* expression, *Hoxc13* expression, *Wnt3* chromatin openess/accessibility and *Wnt3* expression, respectively, on the projected UMAP vector field.

1320 **(b)** Velocities of *Lef1* RNA, *Hoxc13* RNA, *Wnt3* chromatin and *Wnt3* RNA inferred by GraphVelo,
1321 visualized on the projected UMAP vector field.

1322 **(c)** Velocities of *Lef1* RNA, *Hoxc13* RNA, *Wnt3* chromatin and *Wnt3* RNA inferred by MultiVelo,
1323 visualized on the projected UMAP vector field.

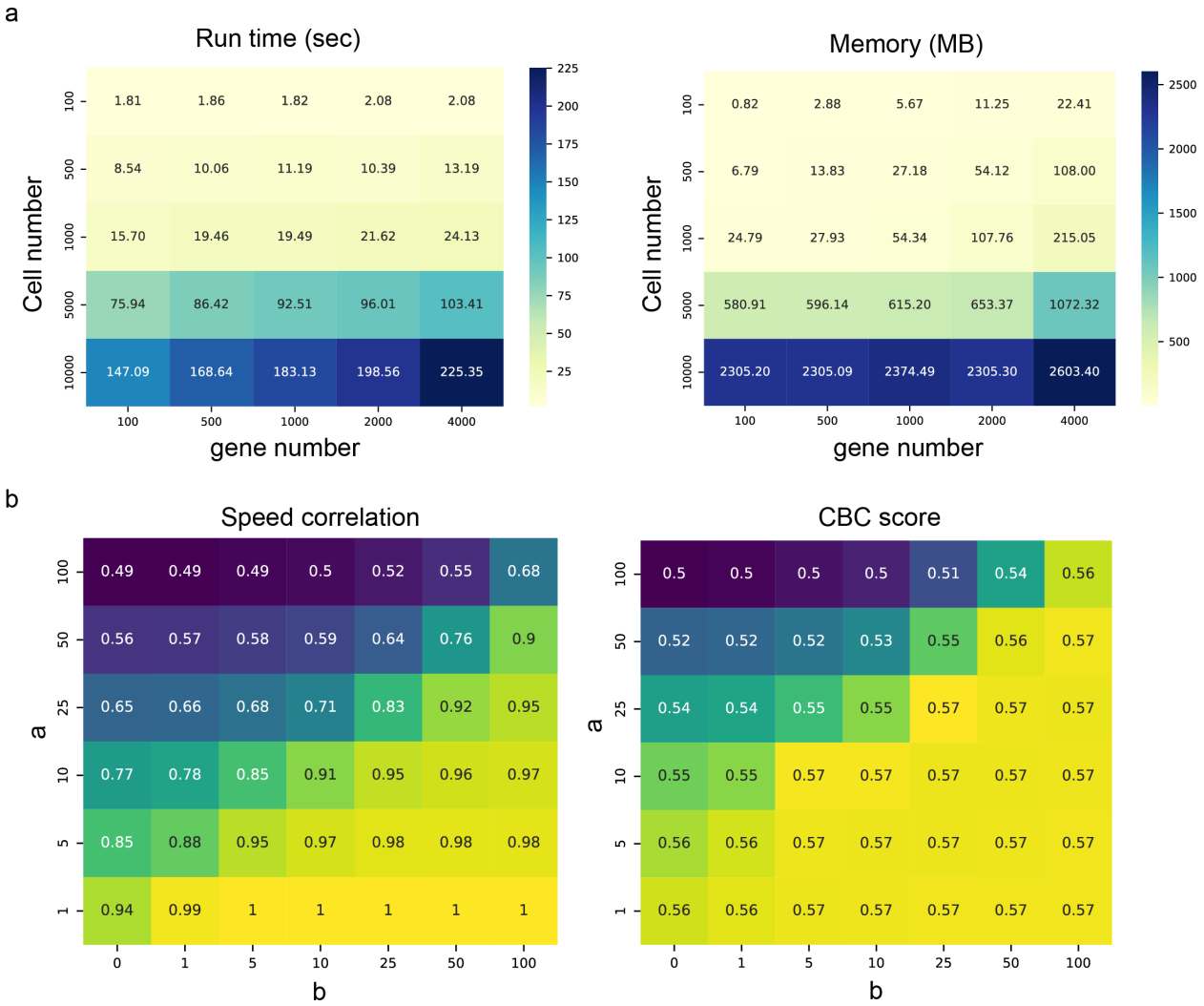
1324 **(d)** Expression of *Hoxc13* gene versus *Lef1* expression. *Lef1* expression primes the activation of
1325 *Hoxc13*.

1326 **(e)** Cell-wise Jacobian analyses of *Lef1-Hoxc13* activation cascade.

1327 **(f)** Jacobian analyses of regulatory interactions between potential PTF *Hoxc13* and i) *Wnt3*
1328 chromatin accessibility or ii) transcription..

1329 **(g)** Jacobian analyses of regulatory interactions between PTF *Lef1* and i) *Wnt3* chromatin
1330 accessibility or ii) transcription.

1331



Extended Data Fig. 16. Sensitivity analysis of GraphVelo on simulated and mouse erythroid dataset.

(a) Runtime (left) and peak memory (right) consumption of GraphVelo with varying numbers of cells and genes on simulated data.

(b) The speed correlation (left) against GraphVelo result reported in Figure. 3 and CBC score (right) with varying hyperparameters *a* and *b*.