POLITECNICO DI TORINO

Corso di Laurea Magistrale
in Ingegneria Matematica

Tesi di Laurea Magistrale

Modelli di RNA velocity per l'inferenza di dinamiche trascrizionali: analisi della stabilità e dell'accuratezza biologica



Relatori
prof. Gianluca Mastrantonio
dott.ssa Elena Sabbioni
firma dei relatori

. . . . . .
. . . . . .

Candidata
Ema-Aldona Danca S347940

firma del candidato

. . . . . .

Anno Accademico 2025-2026

Sommario

A partire da dati scRNA-seq è possibile osservare le quantità di RNA spliced e unspliced per ciascun gene in ciascuna cellula; da queste è possibile inferire l'RNA velocity, ovvero la direzione futura del cambiamento trascrizionale della cellula. Nel corso degli anni si è sviluppata una serie di modelli sempre più complessi per stimare questa quantità.

Il Capitolo 1 introduce il problema, descrive il modello matematico alla base e discute le principali criticità emerse in letteratura riguardo all'affidabilità delle stime prodotte dai metodi attualmente disponibili. Nel Capitolo 2 si descrivono i seguenti metodi di RNA velocity: *scVelo* (framework di riferimento), *cell2fate*, *VeloVI* e *VeloVAE*, considerando anche le varianti *FullVB* e *BrODE*. I metodi sono stati applicati a due dataset di riferimento: l'endocrinogenesi pancreatica e la neurogenesi del giro dentato, che vengono descritti nel Capitolo 3. Il confronto tra questi metodi è contenuto nel Capitolo 4, dove vengono utilizzate alcune metriche quantitative per valutare la correttezza direzionale delle transizioni cellulari, la coerenza locale del campo di velocità e la concordanza tra il tempo latente stimato e l'ordine di sviluppo biologico noto. Si affianca anche l'osservazione diretta dei campi di velocità proiettati nello spazio UMAP, al fine di confrontare la dinamica ricostruita dai diversi metodi con quella biologicamente attesa. Per alcuni metodi viene analizzata anche la stabilità dell'inferenza rispetto all'inizializzazione e agli iperparametri di addestramento.

La tesi fornisce quindi un confronto sistematico tra i diversi approcci, valutando affidabilità, stabilità e coerenza con le dinamiche biologiche attese per comprendere se le criticità osservate derivino da un limite intrinseco del modello cinetico condiviso da tutti gli approcci, oppure da limiti algoritmici legati alle specifiche procedure di inferenza.

Indice

Elenco delle tabelle 5

Elenco delle figure 6

1 Introduzione al problema 7
1.1 Il modello alla base 8

2 Il framework di riferimento e le evoluzioni 11
2.1 *scVelo* 11
2.2 *cell2fate* 20
2.3 *VeloVAE* 24
2.4 *VeloVI* 29

3 Dataset 35

4 Analisi comparativa 41
4.1 Background matematico delle metriche di valutazione 41
4.2 Configurazioni sperimentali 47
4.3 Risultati 49
4.3.1 Endocrinogenesi pancreatica 49
4.3.2 Neurogenesi del giro dentato 58
4.3.3 Confronto tra *scVelo* e *VeloVI* 64
4.4 Conclusioni 69

Elenco delle tabelle

2.1 Confronto tra i modelli di RNA velocity. 34
4.1 Configurazioni sperimentali utilizzate per l'applicazione dei metodi. 48
4.2 Risultati delle metriche di valutazione per i metodi di RNA velocity applicati al dataset del pancreas. 50
4.3 Valori di CBDir ottenuti dai diversi metodi per ciascuna transizione biologica considerata nel dataset del pancreas. 52
4.4 Valori di ICCoh ottenuti dai diversi metodi per ciascun cluster del dataset del pancreas. 53
4.5 Correlazione tra il tempo latente stimato e l'ordine biologico atteso lungo le diverse linee endocrine del dataset del pancreas. 53
4.6 Valori delle metriche di valutazione ottenuti da *cell2fate* sul dataset pancreatico al variare del seed di inizializzazione 54
4.7 Risultati delle metriche di valutazione per i metodi di RNA velocity applicati al dataset del giro dentato. 58
4.8 Valori di CBDir ottenuti dai diversi metodi per ciascuna transizione biologica considerata nel dataset del giro dentato. 60
4.9 Valori di ICCoh ottenuti dai diversi metodi per ciascun cluster del dataset del giro dentato. 61
4.10 Correlazione tra il tempo latente stimato e l'ordine biologico atteso lungo le diverse linee di differenziamento del dataset del giro dentato. 62
4.11 Metriche di valutazione ottenute da *cell2fate* sul dataset del giro dentato al variare del seed di inizializzazione. 62

Elenco delle figure

3.1 Struttura AnnData 36
3.2 Mappa UMAP dello sviluppo epiteliale pancreatico 37
3.3 Mappa UMAP dello sviluppo del giro dentato dell'ippocampo . . . 38
4.1 Confronto delle stime di RNA velocity e del tempo latente ottenute con *scVelo*, *cell2fate*, *VeloVI* e *VeloVAE* (standard, *FullVB*, *BrODE*) sul dataset del pancreas. 51
4.2 Confronto delle RNA velocity e del tempo latente stimati da *cell2fate* sul dataset pancreatico utilizzando differenti seed. 55
4.3 Confronto dei campi di velocità e dei tempi latenti stimati dalle tre varianti di *VeloVAE* con modifiche agli iperparametri sul dataset pancreatico. 57
4.4 Confronto delle stime di RNA velocity e del tempo latente ottenute con *scVelo*, *cell2fate*, *VeloVI* e *VeloVAE* (standard, Full VB, BrODE) sul dataset del giro dentato. 59
4.5 Confronto delle RNA velocity e del tempo latente stimati da *cell2fate* sul dataset del giro dentato utilizzando differenti seed. 63
4.6 Dinamiche inferite da *scVelo* e *VeloVI* nel dataset del pancreas e distribuzione dell'intrinsic e dell'extrinsic uncertainty stimate da *VeloVI*. 66
4.7 *Phase portraits* dei geni *Pcsk2*, *Nnat* e *Cpe* stimati da *scVelo*. . . . 66
4.8 *Phase portraits* dei geni *Pcsk2*, *Nnat* e *Cpe* stimati da *VeloVI*. . . . 67
4.9 Dinamiche inferite da *scVelo* e *VeloVI* nel dataset del giro dentato e distribuzione dell'intrinsic e dell'extrinsic uncertainty stimate da *VeloVI*. 67
4.10 *Phase portraits* dei geni *Ppp3ca*, *Fam155a* e *Hn1* stimati da *scVelo* nel dataset dentate gyrus. 68
4.11 *Phase portraits* dei geni *Ppp3ca*, *Fam155a* e *Hn1* stimati da *VeloVI* nel dataset dentate gyrus. 68

Capitolo 1

Introduzione al problema

Con le nuove tecnologie di trascrittomica, come il sequenziamento dell'RNA a singola cellula (single-cell RNA sequencing, scRNA-seq) è possibile misurare l'espressione genica di singole cellule. Ciò ha permesso quindi di studiare sistemi biologici complessi, come i processi di sviluppo e la differenziazione cellulare. Sono state sviluppate diverse piattaforme sperimentali, tra cui SMART-seq2 e 10x Genomics Chromium, che consentono di analizzare l'espressione genica a livello di singola cellula con diverse profondità di sequenziamento e numero di cellule.

Uno degli obbiettivi principali nell'analisi di questi dati è ricostruire le dinamiche che guidano le transizioni tra diversi stati cellulari. In molti sistemi biologici, infatti, le cellule evolvono lungo traiettorie ben definite: cellule staminali si differenziano in cellule mature, cellule immunitarie si attivano in risposta a infezioni, popolazioni cellulari attraversano differenti stati funzionali durante lo sviluppo... In oncologia, ad esempio, è possibile studiare come le cellule tumorali possano acquisire stati trascrizionali differenti durante la progressione della malattia. Per inferire la direzione futura delle cellule si è introdotto il concetto di RNA velocity, che associa ad ogni cellula un vettore, che fornisce un'indicazione del cambiamento atteso.

Tuttavia, i dati scRNA-seq hanno un limite fondamentale: essi costituiscono osservazioni statiche di cellule che vengono distrutte durante il processo sperimentale. Ogni cellula viene misurata una sola volta e non è quindi possibile seguirne direttamente l'evoluzione temporale. Inoltre, nella realtà, le cellule dei processi biologici che regolano il destino cellulare si muovono lungo una traiettoria di stati continui, mentre i dati sperimentali ci forniscono categorie discrete. Per superare questa limitazione, sono stati sviluppati numerosi metodi computazionali che cercano di inferire una traiettoria temporale di sviluppo, o più in generale una coordinata temporale latente delle cellule, a partire da queste istantanee statiche.

1.1 Il modello alla base

Per descrivere la direzione e la velocità del cambiamento nell'espressione genica nelle singole cellule sono stati sviluppati moltissimi approcci e tutti hanno alla base lo stesso modello in cui si descrive il ciclo di vita dell'RNA messaggero (mRNA) all'interno della cellula. Tale modello si basa sul fatto che questa dinamica possa essere approssimata attraverso tre processi fondamentali, che rappresentano le principali transizioni molecolari che l'mRNA subisce nel tempo.
Il ciclo biologico dell'mRNA può essere schematizzato nei seguenti passaggi:

- **Trascrizione**: il DNA viene copiato in mRNA unspliced, che contiene sia regioni codificanti (esoni) sia regioni non codificanti (introni).
- **Splicing**: l'mRNA unspliced subisce un processo di maturazione in cui gli introni vengono rimossi e le regioni codificanti vengono unite tra loro attraverso il processo di splicing. Il risultato è la formazione dell'RNA maturo, detto spliced.
- **Degradazione**: l'mRNA spliced viene progressivamente degradato nel tempo attraverso meccanismi cellulari di controllo dell'espressione genica.

Per descrivere questa dinamica dal punto di vista matematico, ci si è basati sui principi biologici che regolano questa attività, che costituiscono una rete di reazioni in grado di rappresentare l'evoluzione temporale delle quantità di RNA nelle sue due forme principali. La presenza relativa di queste due componenti permette di distinguere due fasi. Una prima fase di induzione trascrizionale in cui il gene è attivo e vengono prodotte le molecole di RNA unspliced e successivamente di spliced. Quando la trascrizione viene interrotta inizia la fase di repressione in cui non si produce altro RNA unspliced e si converte in spliced solo quello già presente, mentre l'RNA spliced viene degradato. L'ipotesi fondamentale è che la dinamica dell'espressione genica possa essere inferita osservando simultaneamente le quantità di RNA unspliced e spliced misurate nelle singole cellule.

Il processo biologico sottostante viene rappresentato con una chemical reaction network (CRN), in cui la quantità di mRNA varia in modo continuo nel tempo:

$$\varnothing \xrightarrow{\alpha(t)} u(t) \xrightarrow{\beta} s(t) \xrightarrow{\gamma} \varnothing$$

dove $u(t)$ rappresenta la quantità di mRNA unspliced al tempo t ed $s(t)$ la quantità di mRNA spliced, $\alpha(t)$ è il tasso di trascrizione (dipendente dal tempo), β è il tasso di splicing e γ è il tasso di degradazione, che possono essere considerati costanti. Il DNA non viene consumato e si assume possa essere considerato una sorgente

illimitata di produzione di RNA unspliced. Per questo motivo, la produzione di $u(t)$ viene rappresentata come una reazione che ha origine da una sorgente esterna, indicata convenzionalmente con il simbolo ($\varnothing$).

Questo modello fu formalizzato per la prima volta da La Manno et al. [2018] che proposero un sistema di equazioni differenziali (ODE) per descrivere come cambiano nel tempo le quantità $u(t)$ ed $s(t)$:

$$\begin{cases} \dfrac{du_g(t)}{dt} = \alpha_g(t) - \beta_g u_g(t) \\ \dfrac{ds_g(t)}{dt} = \beta_g u_g(t) - \gamma_g s_g(t) \end{cases} \tag{1.1}$$

per ogni gene g, indipendentemente da tutti gli altri geni.

Il tasso di trascrizione $\alpha_g(t)$ descrive la produzione di RNA unspliced e generalmente si assume costante a tratti in modo da descrivere il passaggio tra le fasi di induzione e repressione.

La derivata temporale dell'mRNA spliced è chiamata **RNA velocity** e possiamo indicarla con

$$v_g(t) = \frac{ds_g(t)}{dt},$$

misura la rapidità con cui varia nel tempo la quantità di RNA maturo di un determinato gene e costituisce una stima della direzione futura dello stato trascrizionale della cellula. Quando $v_g(t) > 0$, il sistema si trova generalmente in una fase di induzione trascrizionale, la produzione di RNA spliced supera la sua degradazione; quando invece $v_g(t) < 0$, si osserva generalmente una fase di repressione genica, la degradazione di RNA spliced prevale sulla sua sintesi. Se invece $v_g(t) = 0$ si ha una condizione di equilibrio in cui degradazione e produzione dell'RNA spliced si bilanciano. Una condizione di stato stazionario si ha quando $du_g(t)/dt = 0$ ed anche $ds_g(t)/dt = 0$. È possibile distinguere due stati stazionari: uno attivo (in cui il gene continua a essere trascritto e produzione e degradazione si compensano) e uno inattivo (in cui la trascrizione è interrotta e le quantità di RNA tendono a zero).

Da questa formulazione è nato *velocyto*, che rappresenta il primo metodo di inferenza a questo problema, proposto da La Manno et al. [2018]. In questo primo approccio, l'RNA velocity viene stimata sfruttando la relazione tra le quantità di RNA unspliced e spliced. Si assume che esistano cellule che raggiungono un equilibrio tra trascrizione e degradazione e le quantità $u(t)$ ed $s(t)$ rimangono costanti nel tempo. Queste cellule vengono utilizzate per stimare la relazione tra le due componenti all'equilibrio, tramite una regressione lineare: $u(t) = \gamma s(t)$. La

velocità viene quindi determinata calcolando la distanza rispetto a tale retta come $v(t) = u(t) - \gamma s(t)$. Pur essendo intuitivo e computazionalmente efficiente le ipotesi sono fortemente semplificate, per esempio l'assunzione di linearità e degli stati stazionari, che raramente sono soddisfatte nei sistemi biologici reali.

Successivamente è stato sviluppato *scVelo* da Bergen et al. [2020], che estende il framework originario introducendo un modello dinamico basato sulla massima verosimiglianza, capace di inferire la dinamica trascrizionale completa del sistema: in questo modo è possibile inferire non solo la direzione del cambiamento dell'espressione genica, ma anche assegnare a ciascuna cellula un tempo latente continuo, che rappresenta la sua posizione lungo il processo dinamico.

Negli ultimi anni, si sono affermati altri approcci che cercano di rendere le stime della RNA velocity più robuste. Alcuni metodi introducono diverse formulazioni probabilistiche, altri integrano modelli generativi, mentre altri ancora cercano di catturare meglio la variabilità tra cellule utilizzando diverse assunzioni biofisiche. L'obiettivo comune di questi lavori è quello di rendere la stima della RNA velocity più robusta e più adatta a descrivere sistemi biologici complessi, caratterizzati da biforcazioni e stati transizionali multipli.

Nonostante questi progressi, si evince un problema: in molti casi, infatti, i metodi di RNA velocity producono risultati instabili o biologicamente poco coerenti, che cambiano sensibilmente al variare dei parametri o delle fasi di preprocessing. Ciò ha portato a un dibattito sull'affidabilità biologica delle traiettorie inferite. Infatti i diversi approcci mostrano prestazioni eterogenee a seconda del dataset utilizzato: sebbene alcune volte producano ricostruzioni coerenti delle traiettorie cellulari, altre volte risultano biologicamente poco plausibili. Questi approcci possono risultare inaccurati in contesti particolarmente complessi, come i microambienti tumorali, in cui la presenza di mutazioni può causare, ad esempio, fenomeni di splicing non costante che violano le assunzioni del modello appena descritto.

Questo ha portato a mettere in discussione la modellazione matematica alla base. La stima della RNA velocity è strettamente legata alla parametrizzazione delle equazioni differenziali che descrivono la trascrizione. L'ipotesi che questa dinamica possa essere descritta in modo accurato attraverso un sistema relativamente semplice (basato su tassi costanti o costanti a tratti) potrebbe essere troppo restrittiva per catturare la reale complessità dei sistemi biologici. Di conseguenza, una possibile spiegazione delle criticità osservate è che il limite non sia soltanto algoritmico, ma anche concettuale: ovvero che il modello alla base potrebbe non essere sufficientemente espressivo per rappresentare fedelmente la dinamica cellulare.

Capitolo 2

Il framework di riferimento e le evoluzioni

In questo capitolo viene inizialmente descritto *scVelo*, che resta uno dei riferimenti fondamentali per l'inferenza della dinamica trascrizionale nei dati single-cell. Vengono poi descritti alcuni dei metodi che si sono sviluppati successivamente a *scVelo* e che hanno cercato di estenderlo con diverse assunzioni biofisiche o modificarlo con approcci computazionali diversi. Ciascuno verrà descritto facendo riferimento alla formulazione proposta dagli autori nei lavori originali, evidenziando gli aspetti distintivi rispetto agli altri metodi.

2.1 *scVelo*

In questo framework la stima della RNA velocity può essere ottenuta con tre approcci:

- modello deterministico / a stato stazionario, come *velocyto*: si assume infatti che le cellule raggiungano un equilibrio stazionario. La relazione tra u ed s all'equilibrio viene descritta come una retta, stimata applicando una regressione lineare. Le velocità vengono quantificate come deviazioni rispetto a tale retta. I livelli di equilibrio dell'mRNA vengono stimati selezionando le cellule con le abbondanze di RNA più estreme, cioè quelle nel quantile superiore e in quello inferiore della distribuzione osservata, presunte allo stato stazionario attivo e inattivo.
- modello stocastico: si considerano trascrizione, splicing e degradazione come eventi probabilistici di un processo di Markov a tempo continuo. Si utilizzano, oltre alla media, anche i momenti del secondo ordine per descrivere la variabilità e la covarianza tra le quantità di RNA spliced e unspliced.

- modello dinamico: risolve la dinamica completa per ciascun gene, senza l'assunzione di uno stato stazionario. I parametri cinetici e le variabili latenti cellula-specifiche vengono stimati tramite un algoritmo di Expectation-Maximization (EM).

In seguito prenderemo in considerazione il modello dinamico basato sulla massima verosimiglianza che considera l'intera dinamica trascrizionale gene-specifica. L'analisi si articola in diverse fasi: pre-processing dei dati grezzi, inferenza del modello dinamico e post-processing per l'interpretazione dei risultati.

Preprocessing

Nel framework di *scVelo*, l'analisi è preceduta da una fase di preprocessing dei dati grezzi, che consistono nei conteggi di RNA spliced ed unspliced per ciascun gene e ciascuna cellula. L'obiettivo di questa fase è trasformare tali conteggi discreti in una loro versione approssimativamente continua, che viene utilizzata come input per l'analisi.

Il pre-processing si articola in diversi passaggi. Innanzitutto, viene effettuata una selezione dei geni. Vengono selezionati i geni più variabili (tipicamente 2000) tra quelli che superano la soglia minima di conteggi condivisi tra RNA spliced e unspliced (ad esempio di 20 conteggi). I dati vengono successivamente normalizzati per cellula, dividendo per il numero totale di molecole osservate in ciascuna cellula. Questo serve a correggere differenze tecniche dovute alla diversa profondità di sequenziamento. Si applica una trasformazione logaritmica. Tutti questi passaggi sono implementati nella funzione `scv.pp.filter_and_normalize`.

Una volta effettuati questi passaggi vengono calcolati i momenti delle espressioni geniche, sfruttando la struttura di vicinato tra le cellule. In particolare si costruisce il grafo dei k-nearest neighbors (k-NN), basato sulle distanze euclidee nello spazio PCA calcolate sui conteggi spliced logaritmizzati. Per ogni cellula vengono calcolati la media (momento di primo ordine) e la varianza non centrata (momento di secondo ordine) considerando i 30 vicini più prossimi. Il calcolo dei momenti è implementato nella funzione `scv.pp.moments`, che esegue internamente sia la PCA sia la costruzione del grafo dei vicini.

Inferenza del modello (EM)

Il modello dinamico di *scVelo* è governato dal sistema (1.1), ma per semplicità, nel seguito ometteremo l'indice g. Per ogni cellula c viene introdotto un tempo latente t_c che rappresenta la posizione della cellula lungo la traiettoria dinamica del gene.

In funzione di tale tempo, la cellula può trovarsi in diverse regioni della dinamica trascrizionale, corrispondenti a una fase di induzione, di repressione o a condizioni prossime agli stati stazionari del sistema. Queste fasi possono essere descritte concettualmente come stati trascrizionali discreti k_c anche se sono determinate implicitamente dal valore del tempo latente. Gli stati trascrizionali appartengono all'insieme $k_c \in \{\text{on}, \text{off}, ss^{\text{on}}, ss^{\text{off}}\}$, dove on rappresenta la fase di induzione, ossia l'attivazione della trascrizione, mentre off rappresenta la fase di repressione. Gli stati ss^{on} e ss^{off} corrispondono invece a condizioni di equilibrio (steady-state) del sistema, in cui le quantità di mRNA unspliced e spliced rimangono costanti nel tempo, ossia

$$\frac{du(t)}{dt} = 0, \qquad \frac{ds(t)}{dt} = 0.$$

In tali condizioni si ottengono i valori

$$u^{(k)*} = \frac{\alpha^{(k)}}{\beta}, \qquad s^{(k)*} = \frac{\alpha^{(k)}}{\gamma},$$

dove $\alpha^{(k)}$ assume il valore α^{on} oppure α^{off} a seconda dello stato trascrizionale.

Di conseguenza, il tasso di trascrizione del gene g viene modellato come una funzione costante a tratti:

$$\alpha(t) = \begin{cases} \alpha^{\text{on}} & \text{se } t^{\text{on}} \leq t < t^{\text{off}}, \\ \alpha^{\text{off}} & \text{altrimenti.} \end{cases} \tag{2.1}$$

La fase di repressione è caratterizzata da un livello basale di trascrizione $\alpha^{\text{off}} \geq 0$, mentre nella fase di induzione il tasso di trascrizione è più elevato, cioè $\alpha^{\text{on}} > \alpha^{\text{off}}$. Se t^{on} denota l'istante di inizio della fase di induzione e ω la sua durata, allora l'istante di passaggio alla fase di repressione è dato da $t^{\text{off}} = t^{\text{on}} + \omega$.

Per descrivere le dinamiche negli stati non stazionari osservati nei dati, il modello introduce una traiettoria teorica $\hat{x}(t;\theta) = (\hat{u}(t;\theta), \hat{s}(t;\theta))$ considerando le soluzioni esplicite del sistema (1.1). L'obiettivo dell'inferenza è infatti identificare la traiettoria che meglio approssima i dati osservati; per un dato stato trascrizionale k, le soluzioni per le quantità di mRNA unspliced e spliced sono date da:

$$u^{(k)}(t;\theta) = u_0^{(k)} e^{-\beta\tau^{(k)}} + \frac{\alpha^{(k)}}{\beta}\left(1 - e^{-\beta\tau^{(k)}}\right) \tag{2.2}$$

$$s^{(k)}(t;\theta) = s_0^{(k)} e^{-\gamma\tau^{(k)}} + \frac{\alpha^{(k)}}{\gamma}\left(1 - e^{-\gamma\tau^{(k)}}\right) + \frac{\alpha^{(k)} - \beta u_0^{(k)}}{\gamma - \beta}\left(e^{-\gamma\tau^{(k)}} - e^{-\beta\tau^{(k)}}\right) \tag{2.3}$$

dove $\tau^{(k)} = t - t^{(k)}$ rappresenta il tempo trascorso dall'ultimo cambiamento di stato mentre $u_0^{(k)}$ e $s_0^{(k)}$ sono le condizioni iniziali al tempo $t^{(k)}$.
Per rendere il sistema dinamico ben definito, è necessario specificare tali condizioni iniziali per ciascun regime trascrizionale, che vengono determinate dalla continuità della traiettoria tra stati successivi. Infatti, nel caso di transizione da uno stato k a uno stato k', le condizioni iniziali del nuovo regime sono imposte dalla valutazione della soluzione precedente nel tempo di switching

$$u_0^{(k')} = u^{(k)}\left(t^{(k')}\right), \quad s_0^{(k')} = s^{(k)}\left(t^{(k')}\right).$$

Questa condizione garantisce la continuità della dinamica lungo l'intero processo trascrizionale.

Per ottenere una descrizione esplicita delle fasi di induzione e repressione, è possibile assumere come condizioni iniziali per la fase di induzione i valori $s_0^{\text{on}} = \frac{\alpha^{\text{off}}}{\gamma}$, e $u_0^{\text{on}} = \frac{\alpha^{\text{off}}}{\beta}$. Sostituendo tali condizioni nelle espressioni precedenti (2.2) e (2.3), si ottiene:

$$u^{\text{on}}(t;\theta) = \frac{\alpha^{\text{off}}}{\beta} e^{-\beta\tau^{\text{on}}} + \frac{\alpha^{\text{on}}}{\beta}\left(1 - e^{-\beta\tau^{\text{on}}}\right), \tag{2.4}$$

$$s^{\text{on}}(t;\theta) = \frac{\alpha^{\text{off}}}{\gamma} e^{-\gamma\tau^{\text{on}}} + \frac{\alpha^{\text{on}}}{\gamma}\left(1 - e^{-\gamma\tau^{\text{on}}}\right) + \frac{\alpha^{\text{on}} - \alpha^{\text{off}}}{\gamma - \beta}\left(e^{-\gamma\tau^{\text{on}}} - e^{-\beta\tau^{\text{on}}}\right) \tag{2.5}$$

dove $\tau^{\text{on}} := t - t^{\text{on}}$.
Analogamente, la dinamica nella fase di repressione viene descritta utilizzando come condizioni iniziali i valori raggiunti al termine della fase di induzione, cioè al tempo $t^{\text{off}} = t^{\text{on}} + \omega$

$$u^{\text{off}}(t;\theta) = u^{\text{on}}(t^{\text{on}} + \omega)\, e^{-\beta\tau^{\text{off}}} + \frac{\alpha^{\text{off}}}{\beta}\left(1 - e^{-\beta\tau^{\text{off}}}\right), \tag{2.6}$$

$$s^{\text{off}}(t;\theta) = s^{\text{on}}(t^{\text{on}}+\omega)\, e^{-\gamma\tau^{\text{off}}} + \frac{\alpha^{\text{off}}}{\gamma}\left(1 - e^{-\gamma\tau_{\text{off}}}\right) + \frac{\alpha^{\text{off}} - \beta\, u^{\text{on}}(t^{\text{on}} + \omega)}{\gamma - \beta}\left(e^{-\gamma\tau^{\text{off}}} - e^{-\beta\tau^{\text{off}}}\right) \tag{2.7}$$

dove $\tau^{\text{off}} := t - (t^{\text{on}} + \omega)$.
Quando $t \to \infty$, la dinamica converge allo stato stazionario di repressione

$$ss^{\text{off}} = \lim_{t\to+\infty}\left(u(t;\theta),\, s(t;\theta)\right) = \left(\frac{\alpha^{\text{off}}}{\beta},\, \frac{\alpha^{\text{off}}}{\gamma}\right), \tag{2.8}$$

che corrisponde alle condizioni iniziali del sistema nella fase di induzione.
Se invece la fase on persiste indefinitamente ($\omega = \infty$), il sistema tende allo stato

stazionario di induzione

$$ss^{\text{on}} = \lim_{t\to+\infty} \Big(u^{\text{on}}(t;\theta),\, s^{\text{on}}(t;\theta)\Big) = \left(\frac{\alpha^{\text{on}}}{\beta},\, \frac{\alpha^{\text{on}}}{\gamma}\right). \tag{2.9}$$

Dal punto di vista biologico, ss^{off} rappresenta la condizione in cui il gene si trova in uno stato di bassa o assente trascrizione, mentre ss^{on} corrisponde a uno stato di attivazione trascrizionale più elevata. Durante la dinamica, il sistema viene attratto verso ss^{on} nella fase on e verso ss^{off} nella fase off, tuttavia, se il cambio di stato avviene prima che il sistema abbia avuto il tempo sufficiente per avvicinarsi ai valori asintotici, questi punti non vengono raggiunti.

L'inferenza dei parametri si basa sullo scostamento tra osservazioni sperimentali e traiettoria teorica del modello, assumendo che

$$x_c \sim \mathcal{N}\Big(\hat{x}(t_c;\theta), \sigma^2 I\Big)$$

dove $x_c = (u_c, s_c)$ rappresenta la coppia di quantità unspliced e spliced osservate per la cellula c. Si assume quindi che le osservazioni sperimentali siano generate dalla traiettoria teorica del sistema dinamico affette da rumore gaussiano. Lo scostamento tra dati e modello è quantificato tramite residui di Deming definiti come

$$e_c = \text{sign}\Big(s_c - \hat{s}(t_c)\Big) \cdot \|x_c - \hat{x}(t_c;\theta)\|$$

per ogni cellula c. Si assume che tali residui seguano una distribuzione gaussiana:

$$e_c \sim \mathcal{N}(0, \sigma^2),$$

dove σ^2 è specifico per gene ma costante tra le cellule e le osservazioni sono assunte indipendenti e identicamente distribuite (i.i.d.).

Sotto queste ipotesi, la funzione di verosimiglianza per un singolo gene è

$$L(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2n\sigma^2} \sum_{c=1}^{n} \|x_c - \hat{x}(t_c;\theta)\|^2\right),$$

mentre la corrispondente log-verosimiglianza negativa da minimizzare è

$$\ell(\theta) = \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \cdot \frac{1}{n} \sum_{c=1}^{n} \|x_c - \hat{x}(t_c;\theta)\|^2 .$$

In alternativa, il modello può essere formulato assumendo residui Laplaciani, usando la norma L_1 invece della norma L_2. In entrambi i casi, il comportamento qualitativo del modello rimane invariato.

I parametri del modello $\theta = (\alpha^{(k)}, \beta, \gamma)$ e le variabili latenti vengono stimati tramite un algoritmo di **Expectation-Maximization (EM)**. Date la traiettoria di fase $\hat{x}(t_c;\theta) = (\hat{u}(t_c;\theta), \hat{s}(t_c;\theta))$ e le osservazioni $x_c = (u_c, s_c)$ l'algoritmo EM procede iterativamente alternando i seguenti passi:

- **E-step**: fissati i parametri del modello θ, per ogni osservazione x_c, si stima il tempo latente t_c minimizzando la distanza dalla traiettoria teorica $\hat{x}(t)$

$$t_c = \arg\min_{t_c} \|x_c - \hat{x}(t_c;\theta)\| .$$

- **M-step**: fissati i tempi latenti t_c, si aggiornano i parametri θ massimizzando la verosimiglianza del modello, o, equivalentemente, minimizzando la log-verosimiglianza negativa

$$\theta = \arg\min_{\theta} \, l(\theta) = \arg\min_{\theta} \sum_{c} \|\hat{x}(t_c;\theta) - x_c\|_2^2$$

  a meno di costanti additive e moltiplicative dipendenti da σ^2.

Tuttavia, la stima diretta del tempo latente risulta computazionalmente onerosa, in quanto richiede una minimizzazione, per questo motivo nelle prime fasi viene introdotta una approssimazione del tempo latente.

Sfruttando la struttura analitica delle soluzioni del sistema dinamico è possibile ricostruire il tempo in forma chiusa. Per ciascuna cellula c, il tempo trascorso dall'ingresso nello stato trascrizionale τ_c viene approssimato invertendo le equazioni della dinamica. Per $\beta > \gamma$, la stima del tempo si ottiene a partire dalla combinazione lineare delle quantità spliced e unspliced

$$\tilde{s} = s(t) - \tilde{\beta} u(t), \quad \text{con } \tilde{\beta} = \frac{\beta}{\gamma - \beta}$$

da cui segue

$$\tau_c = -\frac{1}{\gamma} \log\left(\frac{\tilde{s}_c - \tilde{s}_\infty}{\tilde{s}_{0c} - \tilde{s}_\infty} \right)$$

dove $\tilde{s}_\infty = s_\infty - \tilde{\beta} u_\infty$, con s_∞ limite corrispondente allo stato stazionario considerato. Mentre per $\beta \leq \gamma$, si utilizza direttamente la dinamica dell'unspliced

$$\tau_c = -\frac{1}{\beta} \log\left(\frac{u_c - u_\infty}{u_{0c} - u_\infty} \right),$$

dove, analogamente, u_∞ indica il valore asintotico verso cui tende la traiettoria. Questa approssimazione consente di avviare l'algoritmo EM in una regione plausibile dello spazio dei parametri, evitando la necessità di una ricerca globale

non convessa fin dall'inizio. Nella fase M-step, per aggiornare i parametri, viene adottato il metodo del simplesso di Nelder Mead, un metodo di ottimizzazione derivative-free. *ScVelo* fornisce il modulo `scv.tl.recover_dynamics` per stimare le variabili latenti e i tassi.

Post-processing

Per visualizzare e interpretare biologicamente le traiettorie inferite, *scVelo* applica una fase di post-processing in cui le velocità stimate nello spazio ad alta dimensionalità vengono proiettate su uno spazio di embedding a bassa dimensione, tipicamente UMAP. Tale proiezione consente di rappresentare la dinamica del sistema in termini di traiettorie cellulari visualizzabili.

Sia $s_c \in \mathbb{R}^G$ l'espressione genica della cellula c e $\nu_c \in \mathbb{R}^G$ il vettore di velocità associato. La similarità coseno

$$\pi_{cj} = \cos(\delta_{cj}, \nu_c) = \frac{\delta_{cj}^T \nu_c}{\|\delta_{cj}\| \, \|\nu_c\|} \quad \text{con} \quad \delta_{cj} = s_j - s_c, \quad \pi_{cc} = 0$$

misura la coerenza direzionale tra i vettori, indipendentemente dalla magnitudine, varia tra -1 (direzioni opposte), 0 (ortogonali) e 1 (stessa direzione). La matrice di similarità risultante definisce un velocity graph. In tale grafo gli archi rappresentano la compatibilità tra la direzione del vettore di velocità e le possibili transizioni nello spazio delle cellule vicine. L'analisi è ristretta ai vicini in un grafo k-NN per ridurre la complessità. Le similarità vengono poi trasformate in probabilità di transizione tramite un kernel esponenziale, ottenendo una matrice stocastica che descrive una catena di Markov sullo spazio cellulare. Tale matrice, indicata con $\tilde{\pi}$, può essere estratta tramite la funzione `scv.utils.get_transition_matrix`.

Il tempo latente condiviso tra geni viene ottenuto aggregando stime gene-specifiche normalizzate. Le cellule di partenza del processo dette root sono identificate come stati stazionari della catena di Markov

$$\mu^* = \mu^* \tilde{\pi}^T.$$

Il tempo finale di una cellula c rispetto a una root o è stimato come un quantile delle differenze gene-specifiche, mentre il tempo finale per ciascuna cellula viene ottenuto mediando sulle cellule root

$$t_{c,o} = Q_p \left(t_c^g - t_o^g \right), \qquad t_c = \frac{1}{|o|} \sum_o t_{c,o}.$$

Le velocità vengono quindi proiettate su uno spazio a bassa dimensionalità, tipicamente UMAP, mediante la media pesata rispetto alle probabilità di transizione

$$\tilde{\nu}_c = \mathbb{E}_{\tilde{\pi}_c}\left[\tilde{\delta}_c\right] = \sum_{j \neq c} \left(\tilde{\pi}_{cj} - \frac{1}{n}\right)\tilde{\delta}_{cj}$$

dove $\tilde{\delta}_{cj}$, è la direzione degli spostamenti normalizzata, cioè

$$\tilde{\delta}_{cj} = \frac{s_j^{\text{emb}} - s_c^{\text{emb}}}{\left\| s_j^{\text{emb}} - s_c^{\text{emb}} \right\|}$$

dove s_c^{emb} è il vettore delle coordinate della cellula c nello spazio dell'embedding. Questo passaggio è implementato nella funzione `scv.tl.velocity_embedding` e consente di rappresentare la dinamica come un campo vettoriale nel piano dell'embedding.
La visualizzazione delle traiettorie può essere effettuata in diversi modi:

- a livello cellulare con `scv.pl.velocity_embedding`,
- tramite griglie regolari con `scv.pl.velocity_embedding_grid`,
- oppure come linee di flusso con `scv.pl.velocity_embedding_stream`.

Queste rappresentazioni permettono di evidenziare qualitativamente le direzioni principali dei processi di sviluppo cellulare. Inoltre, è possibile identificare stati iniziali e finali delle traiettorie cellulari tramite la funzione `scv.tl.terminal_states`, che analizza la catena di Markov associata alla matrice di transizione.

Tuttavia, l'interpretazione biologica non deve basarsi esclusivamente sulle velocità proiettate, ma deve essere supportata dall'analisi delle dinamiche geniche a livello individuale. Per questo motivo, *scVelo* utilizza i *Phase portraits*, che rappresentano la relazione tra RNA unspliced e spliced per ciascun gene, visualizzati tramite le funzioni: `scv.pl.velocity` e `scv.pl.scatter`, che permettono di analizzare singoli geni e verificare la coerenza tra dinamica genica e direzione delle velocità inferite.

Evoluzioni dei modelli di RNA velocity

Nonostante il successo di *scVelo*, alcuni risultati ottenuti mettono in dubbio l'affidabilità dell'RNA velocity. Gli stessi autori affermano che alcune stime risultano fortemente distorte, con traiettorie cellulari che contraddicono le aspettative biologiche, Bergen et al. [2021]. Tanto che ci si chiede quanto questi strumenti siano affidabili in applicazioni in cui queste non sono note. Sebbene *scVelo* rappresenti un significativo avanzamento rispetto ai metodi precedenti, il modello presenta

ancora alcune limitazioni rilevanti. In primo luogo, l'inferenza è effettuata a livello di singolo gene, il che implica che il tempo latente venga inizialmente stimato separatamente per ciascun gene, questo può rendere difficile l'interpretazione globale del processo biologico. Un ulteriore aspetto critico riguarda la forte dipendenza del modello dalle fasi di preprocessing dei dati. Le stime di RNA velocity possono variare significativamente in funzione delle scelte effettuate nelle fasi preliminari, come la normalizzazione, il filtraggio dei geni, la selezione delle cellule e i metodi utilizzati per quantificare le molecole di RNA spliced e unspliced. L'assunzione di normalità dei dati è spesso irrealistica per dati scRNA-seq, caratterizzati da basse numerosità. Differenze in queste procedure possono portare a risultati non coerenti tra diverse analisi dello stesso dataset. Inoltre, *scVelo* ha bisogno anche di una fase di post-processing; in particolare, utilizza tecniche di proiezione che possono distorcere i risultati, Zheng et al. [2023]. Questo processo tende a privilegiare la struttura locale del grafo dei vicini più prossimi (k-NN), facendo sì che le velocità riflettano quelle dei vicini piuttosto che nuove traiettorie. Di conseguenza la qualità delle stime dipende fortemente dalla costruzione del grafo k-NN: in presenza di rumore, eventuali errori nella sua definizione possono propagarsi. Un limite rilevante è l'assenza di una quantificazione esplicita dell'incertezza, in quanto si basa esclusivamente su stime puntuali, riducendo quindi la fiducia nelle sue conclusioni.

Queste limitazioni hanno motivato lo sviluppo di approcci più sofisticati, che cercano di rilassare una o più delle assunzioni critiche del framework originale. In particolare, tali modelli mirano a migliorare l'accuratezza delle stime, la capacità di descrivere dinamiche biologiche complesse e la robustezza delle inferenze. Per affrontare queste limitazioni, sono stati introdotti modelli basati su approcci generativi e su inferenza bayesiana.

Tra questi troviamo *cell2fate*, Aivazidis et al. [2025], un modello bayesiano che linearizza le dinamiche del tasso di trascrizione decomponendolo in un numero ridotto di moduli indipendenti, ciascuno con tempi di accensione/spegnimento specifici per cellula che permette di rappresentare dinamiche complesse come combinazioni di processi più semplici. Questo permette di catturare dinamiche multistep complesse in modo analiticamente trattabile e biologicamente interpretabile. Un aspetto distintivo di *cell2fate* è la capacità di stimare, per ciascuna cellula, una distribuzione di probabilità sui possibili destini cellulari.

VeloVAE definito da Gu et al. [2022] invece utilizza reti neurali per ricostruire la dinamica dell'espressione genica, modellizzando esplicitamente l'emergere di linee cellulari multiple. Il suo funzionamento si basa sulla teoria dei variational autoencoders (VAEs): l'encoder, il modello di inferenza, è una rete neurale che

analizza i conteggi osservati e stima le probabilità a posteriori del tempo latente globale, uno stato cellulare latente continuo. Il modello generativo è invece il decoder, rappresentato da una seconda rete neurale che prende lo stato e calcola i parametri cinetici della trascrizione.

VeloVI è un modello deep generative sviluppato da Gayoso et al. [2024], sempre basato sui VAEs che combina il framework dei modelli generativi profondi con una formulazione bayesiana che permette di stimare esplicitamente l'incertezza associata alla RNA velocity. In questo caso, il modello apprende una rappresentazione latente delle cellule che integra informazione temporale e stato trascrizionale, producendo una distribuzione a posteriori delle velocità per ciascuna cellula e gene.

2.2 *cell2fate*

cell2fate è un modello di RNA velocity completamente bayesiano, sviluppato per superare la difficoltà nel descrivere dinamiche trascrizionali complesse, in cui i tassi di trascrizione $\alpha_g(t)$, dipendenti dal tempo, possono essere complessi e non lineari. L'idea centrale del metodo è sostituire la descrizione del tasso di trascrizione di ciascun gene con una decomposizione in un numero ridotto di moduli dinamici, condivisi tra geni.

Come nei modelli classici, la dinamica di un gene è descritta dai processi di trascrizione, splicing e degradazione. Tuttavia, mentre in *scVelo* il tasso di trascrizione $\alpha_g(t)$ viene modellato come funzione costante a tratti con un singolo passaggio tra stato on e stato off, in *cell2fate* esso viene rappresentato come somma di più componenti:

$$\alpha_g(t) = \sum_{m=1}^{M} \alpha_{mg}(t), \tag{2.10}$$

dove $m = 1, \ldots, M$ indica il modulo. Dal punto di vista biologico, un modulo può rappresentare, per esempio, un programma di proliferazione, un programma di differenziamento oppure una fase tardiva di maturazione cellulare. Ogni modulo contribuisce quindi al tasso di trascrizione del gene g con una dinamica propria, permettendo di rappresentare pattern più complessi rispetto alla semplice alternanza induzione/repressione. Pertanto la derivata temporale del tasso di trascrizione viene modellata come somma di termini linearizzabili:

$$\frac{d\alpha_g(t)}{dt} = \sum_{m=1}^{M} \frac{d\alpha_{mg}(t)}{dt}. \tag{2.11}$$

Per ciascun modulo m, la variazione del contributo trascrizionale è descritta da

$$\frac{d\alpha_{mg}(t)}{dt} = \lambda_{mi}\Big(\hat{\alpha}_{mgi} - \alpha_{mg}(t)\Big), \tag{2.12}$$

dove:

- $\hat{\alpha}_{mgi}$ è il tasso di trascrizione target del modulo m per il gene g;
- λ_{mi} controlla la velocità con cui il modulo raggiunge il suo valore target;
- $i \in \{\text{on}, \text{off}\}$ indica lo stato del modulo.

Questa formulazione implica che ogni modulo si accenda e si spenga gradualmente, invece di cambiare istantaneamente stato. I tempi di accensione e spegnimento del modulo m sono indicati con t_m^{on} e t_m^{off}, e vengono confrontati con un tempo latente cellula-specifico t_c. In altre parole, ogni cellula c è collocata lungo una coordinata temporale continua, e la sua posizione rispetto a t_m^{on} e t_m^{off} determina se il modulo è in fase di induzione oppure in fase di repressione.

La decomposizione (2.10)–(2.12) rende il modello analiticamente trattabile, poiché le equazioni differenziali risultanti possono essere risolte in forma chiusa; inoltre introduce una struttura interpretabile, nella quale i moduli possono essere letti come programmi biologici condivisi tra più geni. Sostituendo la decomposizione modulare nelle equazioni (1.1) della RNA velocity, per ciascun gene g si ottiene:

$$\begin{cases} \dfrac{du_g(t)}{dt} = \displaystyle\sum_{m=1}^{M} \frac{du_{mg}(t)}{dt} = \sum_{m=1}^{M} \Big(\alpha_{mg}(t) - \beta_g u_{mg}(t)\Big), \\ \dfrac{ds_g(t)}{dt} = \displaystyle\sum_{m=1}^{M} \frac{ds_{mg}(t)}{dt} = \sum_{m=1}^{M} \Big(\beta_g u_{mg}(t) - \gamma_g s_{mg}(t)\Big). \end{cases} \tag{2.13}$$

dove $u_g(t)$ e $s_g(t)$ rappresentano rispettivamente le quantità di mRNA unspliced e spliced del gene g, mentre β_g e γ_g sono i tassi gene-specifici di splicing e degradazione. In modo analogo, $u_{mg}(t)$ e $s_{mg}(t)$ rappresentano il contributo del modulo m alle quantità unspliced e spliced del gene g. Si ha quindi

$$u_g(t) = \sum_{m=1}^{M} u_{mg}(t), \qquad s_g(t) = \sum_{m=1}^{M} s_{mg}(t).$$

Da qui possiamo definire anche

$$v_{mg} = \frac{ds_{mg}}{dt},$$

come la velocità per ciascun modulo. La velocità totale sarà la somma delle velocità di ciascun modulo.

A differenza di *scVelo*, dove la dinamica di ciascun gene viene trattata in modo sostanzialmente indipendente, in *cell2fate* i geni sono accoppiati indirettamente attraverso i moduli condivisi. I moduli di *cell2fate* possono essere interpretati come programmi di espressione genica attivati sequenzialmente nel tempo. In questo senso, il modello può essere interpretato anche come un modello a *membership* mista: l'espressione dinamica di ciascun gene non è attribuita a un unico stato discreto, ma a una combinazione continua di più moduli. La successione temporale dei moduli permette quindi di scomporre una traiettoria complessa in finestre trascrizionali distinte.

Uno degli output principali del modello è il tempo latente specifico per cellula t_c, che ordina le cellule lungo una traiettoria dinamica comune. A differenza di metodi *gene-wise*, il tempo in *cell2fate* viene inferito in modo congiunto usando tutti i geni e tutti i moduli, risultando quindi più stabile e interpretabile a livello globale. Dato il tempo t_c, per ciascun modulo si può definire uno stato dinamico: off, se il modulo non è ancora attivo; induzione, se il modulo si sta accendendo; on, se il modulo è pienamente attivo; repressione, se il modulo si sta spegnendo.

Un'altra differenza è che *cell2fate* non lavora su dati preprocessati e approssimativamente gaussiani, ma direttamente sui **conteggi grezzi** di RNA spliced e unspliced del sequenziamento. Questo è reso possibile introducendo un modello di misura esplicito che separa il segnale biologico dagli effetti tecnici, gestendo la variabilità tecnica incorporando diversi fattori. Sia $j \in \{u, s\}$ indice che identifica la maturità dell'RNA e sia $x^B_{cgj} = (u_{cg}, s_{cg})$ il vettore dei valori attesi biologici per la cellula c e il gene g, dove u_{cg} e s_{cg} rappresentano rispettivamente i conteggi attesi di RNA unspliced e spliced in assenza di effetti tecnici. Per tenere conto degli errori di misura, il modello introduce quindi i valori attesi osservati x^M_{cgj}:

$$x^M_{cgj} = l_{cgj}\Big(H_{ce}s_{egj} + x^B_{cgj}\Big), \tag{2.14}$$

dove:

- l_{cgj} rappresenta l'efficienza di rilevazione, che tiene conto di effetti come profondità di sequenziamento e differenze di quantificazione;
- H_{ce} è una codifica *one-hot* dell'assegnazione della cellula c al batch sperimentale e;
- s_{egj} modella la contaminazione dovuta all'RNA ambientale.

Con queste scelte il modello è in grado di rimuovere l'influenza dell'RNA ambientale e le distorsioni tecniche tra esperimenti diversi.

I conteggi osservati X_{cgj} vengono modellati tramite due distribuzioni binomiali negative indipendenti (una per la componente unspliced e una per quella spliced)

$$X_{cgj} \sim \mathrm{NB}\Big(\mu = x^{M}_{cgj},\, a_{gj}\Big), \tag{2.15}$$

dove a_{gj} è un parametro di overdispersione specifico per gene. L'uso di questa distribuzione modella il rumore fisiologico "extra" come i *transcriptional burst*, fenomeni esplosivi della sintesi genica. Questa scelta è più realistica rispetto all'assunzione gaussiana usata in *scVelo*, poiché i dati scRNA-seq sono sparsi, discreti e rumorosi.

Tutti i parametri del modello vengono dotati di distribuzioni a priori gerarchiche e stimati in un quadro completamente bayesiano. Questo implica che il modello non restituisce solo stime puntuali, ma distribuzioni a posteriori dei parametri. Di conseguenza, oltre alla stima del tempo cellulare e della RNA velocity, *cell2fate* fornisce anche una quantificazione esplicita dell'incertezza. Per esempio, per ogni cellula è possibile ottenere non solo una stima di t_c, ma anche la deviazione standard della sua distribuzione a posteriori, che misura quanto il modello sia sicuro della posizione temporale assegnata a quella cellula. Questa è una differenza sostanziale rispetto a *scVelo*, che si basa su stime puntuali.

Siano X_{cgj} i dati osservati e $\theta_{cgj} = (l_{cgj}, H_{ce}, s_{egj}, t_c, \beta_g, A_{mg\mathrm{on}}, \lambda_{mi}, T_{mi})$ l'insieme di tutti i parametri. Per stimare la distribuzione a posteriori dei parametri del modello applichiamo il teorema di Bayes

$$P(\theta_{cgj} \mid X_{cgj}) \propto P(X_{cgj} \mid \theta_{cgj})\, P(\theta_{cgj}),$$

dove $P(X_{cgj} \mid \theta_{cgj})$ è la full likelihood e $P(\theta_{cgj})$ rappresenta le distribuzioni a priori.

Le prior sono scelte in modo gerarchico, riflettendo la struttura del problema. Ad esempio, i tempi cellulari t_c sono modellati tramite una distribuzione normale con media e varianza a loro volta estratte da distribuzioni Gamma, mentre i tassi di splicing e degradazione sono assegnati a distribuzioni Gamma con iperparametri anch'essi distribuiti come una Gamma.

Siccome la distribuzione a posteriori non è calcolabile in forma chiusa a causa della complessità del modello e dell'enorme quantità di parametri, *cell2fate* utilizza la *stochastic variational inference* (SVI), implementata in Pyro, per approssimarla. Viene introdotta una distribuzione approssimante $q(\theta)$, appartenente a una famiglia semplice (ad esempio gaussiane trasformate), e si ottimizzano i parametri in modo che sia il più possibile vicina alla vera distribuzione a posteriori. Questa vicinanza è misurata tramite la divergenza di Kullback–Leibler:

$$\mathrm{KL}\Big(q(\theta) \,\|\, P(\theta \mid X)\Big).$$

Poiché la distribuzione a posteriori è sconosciuta, la minimizzazione della KL viene effettuata massimizzando la cosiddetta *Evidence Lower Bound* (ELBO):

$$\text{ELBO} = \mathbb{E}_q\Big[\log P(X \mid \theta)\Big] - \text{KL}\Big(q(\theta) \,\|\, P(\theta)\Big). \tag{2.16}$$

Il primo termine misura quanto bene il modello spiega i dati, mentre il secondo penalizza deviazioni eccessive rispetto alle prior, introducendo una forma di regolarizzazione bayesiana. L'ottimizzazione dell'ELBO viene effettuata tramite stochastic gradient descent (Adam optimizer), per 500 iterazioni, learning rate 0.01 e batch size di 1000 cellule come default.

Per migliorare l'approssimazione variazionale, *cell2fate* utilizza *AutoHierarchical NormalMessenger* di *Pyro*. In particolare, ogni parametro è modellato tramite una distribuzione normale e poi viene trasformato nel dominio corretto

$$\theta = \text{transform}\Big(\mathcal{N}(\mu, \sigma)\Big),$$

dove μ è la media totale della distribuzione variazionale, σ la deviazione standard indipendente, transform garantisce che il parametro appartenga al dominio corretto (ad esempio valori positivi per distribuzioni Gamma). La media della distribuzione variazionale non è indipendente, ma dipende dalla prior gerarchica

$$\mu = \text{loc} + \text{weight} \cdot \text{transform}^{-1}(\hat{\mu}),$$

dove $\hat{\mu}$ rappresenta la media della prior, loc è la componente di media indipendente. In questo modo, la stima dei parametri è guidata sia dai dati sia dalla struttura gerarchica del modello.

Nel complesso, *cell2fate* presenta diversi vantaggi: permette di descrivere dinamiche trascrizionali complesse e multi-step senza ricorrere a forti semplificazioni oppure a soluzioni puramente numeriche. In secondo luogo, fornisce una struttura modulare interpretabile, che consente di associare i pattern dinamici a programmi biologici concreti. Inoltre, il modello lavora direttamente sui conteggi grezzi, integrando effetti tecnici come batch, differenze di detection e RNA ambientale. Infine, grazie alla formulazione bayesiana, fornisce una quantificazione esplicita dell'incertezza, elemento fondamentale per valutare l'affidabilità delle traiettorie inferite.

2.3 *VeloVAE*

Questo modello implementa i concetti biologici di trascrizione, splicing e degradazione attraverso un modello generativo che utilizza un *Variational AutoEncoder*

(VAE) all'interno di un framework bayesiano. Ogni cellula c viene caratterizzata da due variabili latenti: un tempo cellulare globale t_c, condiviso da tutti i geni all'interno della stessa cellula, e uno stato cellulare latente continuo $\mathbf{z}_c$, che descrive la posizione della cellula su una varietà a bassa dimensionalità associata al processo di sviluppo. Invece di assumere che il tasso di trascrizione sia costante a tratti e dipendente solo da uno stato discreto on/off, *VeloVAE* introduce un tasso di trascrizione specifico per cellula, che dipende dallo stato cellulare latente.

Come nei modelli classici di RNA velocity la dinamica dell'espressione genica viene descritta a partire dalle quantità di RNA unspliced e spliced; viene effettuata una prima fase di preprocessing, che in questo framework è identica a quella di *scVelo*. Tuttavia, invece di avere

$$\alpha(t) \in \{\alpha^{\text{on}}, \alpha^{\text{off}}\},$$

VeloVAE assume che, per ciascun gene, il tasso effettivo di trascrizione sia modulato da un fattore relativo $\rho_c \in [0,1]$, specifico della cellula, che rappresenta il livello di attivazione trascrizionale nella cellula c:

$$\alpha' = \rho_c \alpha \tag{2.17}$$

in cui $\rho_c = 1$ corrisponde all'induzione e $\rho_c = 0$ alla repressione. In questo modo il modello non impone più una dinamica con un'unica fase di induzione seguita da una fase di repressione, ma consente variazioni continue della trascrizione durante lo sviluppo.

Il parametro ρ_c non viene stimato indipendentemente per ogni cellula, ma viene modellato come funzione dello stato cellulare latente:

$$\rho_c = g(\mathbf{z}_c), \tag{2.18}$$

dove $g(\cdot)$ è una rete neurale. In questo modo si introduce una variabilità cellulare che dipende dallo stato latente: cellule vicine tendono ad avere tassi di trascrizione simili, mentre cellule appartenenti a rami diversi della differenziazione possono mostrare dinamiche trascrizionali diverse. Questa formulazione permette di descrivere processi di differenziamento ramificati, che risultano difficili da rappresentare con una successione rigida di fasi on/off.

Poiché ρ_c è considerato costante nel tempo per la singola cellula, ma variabile da cellula a cellula, il sistema (1.1) ammette ancora una soluzione analitica.

Dal punto di vista probabilistico, *VeloVAE* assume che ogni cellula sia generata da un processo latente. Per la cellula c, il tempo latente t_c e lo stato cellulare $\mathbf{z}_c$ vengono trattati come variabili casuali

$$t_c \sim \mathcal{N}(t_0, \sigma_0^2), \qquad \mathbf{z}_c \sim \mathcal{N}(0, I_d), \tag{2.19}$$

dove t_0 e σ_0^2 rappresentano rispettivamente la media e la varianza della distribuzione a priori del tempo latente. Analogamente, lo stato cellulare latente è modellato mediante una distribuzione normale multivariata a media nulla e matrice di covarianza I_d, dove d è la dimensione dello spazio latente. Nell'implementazione proposta dagli autori, lo spazio latente ha dimensione $d = 5$, ma può arrivare anche a 10 per dataset di grandi dimensioni, per cui ogni cellula è rappresentata da un vettore $\mathbf{z}_c \in \mathbb{R}^d$.

I dati osservati, ovvero il vettore concatenato dei conteggi osservati di RNA unspliced e spliced della cellula $x_c = (u_c, s_c) \in \mathbb{R}^{2G}$, sono generati secondo

$$x_c \sim \mathcal{N}\Big(F(t_c; \theta), \Sigma\Big), \tag{2.20}$$

dove $F(t_c; \theta)$ rappresenta la soluzione del sistema cinetico per tutti i geni e Σ è una matrice di covarianza, che si assume diagonale, che modella il rumore di osservazione. In altre parole, il modello assume che i conteggi osservati siano una versione rumorosa della traiettoria predetta dalle ODE.

L'inferenza viene effettuata tramite un variational autoencoder bayesiano, poiché la distribuzione a posteriori non è calcolabile in forma chiusa. Si introduce una distribuzione variazionale approssimante $q_\phi(t_c, \mathbf{z}_c \mid x_c)$, parametrizzata da una rete neurale con parametri ϕ. Questa rete costituisce l'encoder del VAE: un Multilayer Perceptron (MLP) con 2 layers nascosti, riceve in input il vettore concatenato x_c e restituisce i parametri della loro distribuzione a posteriori approssimata, ossia la media e la varianza del tempo latente e, analogamente, il vettore delle medie e la matrice di covarianza dello stato cellulare latente.

Dalle distribuzioni apprese vengono successivamente campionati il tempo cellulare latente t_c e il vettore latente $\mathbf{z}_c$, che costituiscono l'input del decoder.

Il decoder, invece, non è una rete neurale completamente libera come nei VAE classici: esso combina una rete neurale che implementa la funzione $g(\mathbf{z}_c)$, che determina il fattore di modulazione trascrizionale ρ_c, con la soluzione analitica delle ODE. Il decoder non apprende quindi una funzione arbitraria di ricostruzione, ma genera i conteggi attesi di RNA unspliced e spliced imponendo che essi rispettino le equazioni cinetiche del modello. In questo senso, *VeloVAE* è un modello generativo strutturato, in cui la ricostruzione dei dati avviene imponendo esplicitamente una dinamica biologicamente plausibile.

Una volta che il decoder ha ricostruito i conteggi di RNA, il modello confronta tali valori con le osservazioni. La funzione obiettivo ottimizzata durante il training è la *Evidence Lower Bound* (ELBO)

$$\text{ELBO} = \sum_{c=1}^{N} \Big[\mathbb{E}_{q_\phi(t_c, \mathbf{z}_c \mid x_c)} \left[\log p(x_c \mid t_c, \mathbf{z}_c)\right] - D_{\text{KL}} \left(q_\phi(t_c, \mathbf{z}_c \mid x_c) \,\|\, p(t_c, \mathbf{z}_c)\right)\Big],$$

mediante mini-batch stochastic gradient descent.

L'addestramento del modello si articola in una prima fase nella quale vengono inferiti il tempo cellulare latente, lo stato cellulare e i parametri cinetici delle ODE massimizzando la funzione ELBO mediante ottimizzazione stocastica a mini-batch. La soluzione delle equazioni cinetiche dipende anche dalle condizioni iniziali (u_{0c}, s_{0c}) associate a ciascuna cellula. Poiché tali quantità non sono direttamente osservabili, durante la prima fase esse vengono inizializzate in modo approssimativo. Una volta ottenuta una prima stima del tempo cellulare latente, il modello utilizza l'ordinamento temporale inferito per aggiornare queste condizioni iniziali. In particolare, utilizzando il tempo latente stimato, per ciascuna cellula vengono individuate cellule vicine collocate in una finestra temporale precedente; le relative abbondanze spliced e unspliced vengono quindi utilizzate per aggiornare le condizioni iniziali. Il modello viene successivamente riaddestrato sulla base delle nuove condizioni iniziali, in una fase indicata nell'implementazione come *post-training*. La procedura può essere ripetuta fino alla convergenza, consentendo di affinare progressivamente la ricostruzione delle dinamiche cellulari.

VeloVAE consente anche di incorporare informazione temporale esterna. Se per ogni cellula è noto un *capture time*, cioè il timepoint sperimentale in cui la cellula è stata sequenziata, questo può essere utilizzato come prior informativo sul tempo latente

$$t_c \sim \mathcal{N}(t_{\text{cap},c}, \sigma_0^2). \tag{2.21}$$

Ciò rende possibile ancorare il tempo latente a una scala temporale reale, espressa per esempio in giorni o ore, e interpretare di conseguenza anche i parametri cinetici in unità fisiche.

Viene proposta un'estensione del modello che consente di quantificare l'incertezza dei tassi delle ODE, chiamata *FullVB*. Sia $\theta = (\alpha, \beta, \gamma)$ l'insieme dei parametri cinetici e sia

$$p_\lambda(\theta) = \prod_{i=1}^{3} \text{LogNormal}(\theta_i \mid \mu_i, \sigma_i^2),$$

la distribuzione a priori, in cui ciascun parametro è distribuito secondo una log-normale. Di conseguenza, la log-verosimiglianza marginale dei dati ammette come limite inferiore

$$\log p_\lambda(X) \geq \mathbb{E}_{q_\phi(\theta)} \left[\text{ELBO}(X; \theta)\right] - D_{\text{KL}} \left(q_\phi(\theta) \,\|\, p_\lambda(\theta)\right). \tag{2.22}$$

A differenza di *scVelo*, che usa un'ottimizzazione con l'algoritmo di Nelder Mead dei parametri cinetici per ogni gene, *VeloVAE* utilizza mini-batch stochastic gradient descent per addestrare simultaneamente reti neurali e parametri ODE. La distribuzione a posteriori di ciascun parametro $\theta = (\alpha, \beta, \gamma)$ viene approssimata

da $q_\phi(\theta \mid x)$. Tale formulazione aiuta inoltre a prevenire il cosiddetto *vanishing velocity problem*: senza regolarizzazione, l'ottimizzazione potrebbe spingere i tassi cinetici verso valori troppo grandi, ottenendo una buona ricostruzione dei dati ma una velocity prossima a zero per quasi tutte le cellule. L'introduzione di prior sui parametri e della penalizzazione KL impedisce questa degenerazione.

Dal punto di vista biologico, il vantaggio principale di *VeloVAE* è che il tempo latente non viene più inferito separatamente per ciascun gene, ma in modo congiunto usando tutte le informazioni trascrizionali della cellula. Questo produce una coordinata temporale globale più coerente e interpretabile. Inoltre, la presenza dello stato cellulare continuo permette di rappresentare l'emergere di molteplici linee cellulari all'interno di uno stesso modello, senza imporre una struttura ad albero discreta a priori.

Per aumentare ulteriormente l'interpretabilità biologica, gli autori introducono anche un'estensione chiamata branching ODE, *BrODE*, nella quale i tipi cellulari discreti e le relazioni progenitore–discendente vengono inferiti a partire dai risultati di *VeloVAE*. Nella formulazione si assume che:

1. Ogni cellula appartiene a un solo tipo $y \in y_1, \ldots, y_k$.
2. Esiste almeno un tipo staminale (root).
3. Ogni tipo non staminale ha un solo progenitore $par(y)$.

Questa struttura definisce un grafo diretto: $G = (V, E)$: l'arco (u, v) indica che u si differenzia in v. Sotto queste ipotesi, ciascun tipo cellulare y è caratterizzato da parametri cinetici specifici $(\alpha_y, \beta_y, \gamma_y)$ e pertanto ha una ODE unica. Le condizioni iniziali associate a ciascun tipo cellulare sono definite in modo ricorsivo, dipendono dallo stato del tipo progenitore al tempo di emergenza $t_0(y)$ e sono definite come:

$$u_0(y) = \begin{cases} u_{init} & \text{se } y \text{ è una cellula root} \\ u(t_0(y), par(y)) & \text{altrimenti} \end{cases} \tag{2.23}$$

$$s_0(y) = \begin{cases} s_{init} & \text{se } y \text{ è una cellula root} \\ s(t_0(y), par(y)) & \text{altrimenti} \end{cases} \tag{2.24}$$

e questo porta a trovare le soluzioni (2.3) e (2.2) ricorsivamente.

Per applicare il modello di branching ODE è necessario conoscere il grafo delle transizioni cellulari, tale informazione non è generalmente disponibile, pertanto è necessario inferirla dai dati. Le cellule vengono raggruppate mediante Leiden clustering, per ogni coppia di tipi cellulari viene stimata una probabilità di transizione

sulla base dei vicini più prossimi nello spazio latente. Infine si usa l'algoritmo di Edmonds per costruire l'arborescenza di peso massimo.

Assumendo che il tempo cellulare sia già stato inferito tramite VeloVAE, il modello branching ODE viene formulato come un problema di regressione e sostituito al decoder di VeloVAE. I parametri cinetici $\theta = (\alpha_y, \beta_y, \gamma_y)$ vengono stimati massimizzando la likelihood gaussiana, equivalente alla minimizzazione di:

$$\min_{\theta} \sum_{c=1}^{n} (\hat{x}(t_c, \theta) - x_c)^T \Sigma (\hat{x}(t_c, \theta) - x_c) \tag{2.25}$$

dove x_c è il livello di espressione genica osservato e $\hat{x}(t_c, \theta)$ la predizione del modello. L'ottimizzazione viene effettuata mediante mini-batch stochastic gradient descent utilizzando Adam optimizer.

Sebbene questa estensione sia più interpretabile, il modello principale resta quello continuo definito precedentemente, che rappresenta il nucleo innovativo del metodo. Nel complesso, *VeloVAE* rappresenta un'estensione significativa dei modelli classici di RNA velocity. Rispetto a *scVelo*, conserva l'interpretabilità delle equazioni cinetiche ma sostituisce la dinamica discreta on/off con una modulazione continua guidata dallo stato cellulare latente. Rispetto a *cell2fate*, non introduce una decomposizione modulare del tasso di trascrizione, ma punta a modellare direttamente le variazioni trascrizionali come funzione di una variabile latente continua.

2.4 *VeloVI*

VeloVI è un modello di deep learning, introdotto da Gayoso et al. [2024], progettato per migliorare la stima della RNA velocity, che riformula il problema dell'inferenza delle dinamiche trascrizionali all'interno di un framework bayesiano basato sempre su *variational inference* per approssimare distribuzioni a posteriori complesse.

Come nel modello dinamico di *scVelo*, la dinamica di ciascun gene viene descritta, dal sistema di equazioni differenziali (1.1); in particolare, si basa sull'idea che le quantità di RNA unspliced $u(t)$ e spliced $s(t)$ presenti in ogni gene di una cellula siano generate in funzione dei tassi cinetici. Si assume anche l'esistenza di due variabili latenti: un tempo e uno stato trascrizionale, specifici per gene e cellula. In aggiunta, il modello presume che i tempi latenti siano collegati tramite una variabile latente a bassa dimensione ($d = 10$)

$$z_c \sim \mathcal{N}(0, I_d), \tag{2.26}$$

per ogni cellula c, che rappresenta lo stato fenotipico complessivo della cellula e costituisce il collegamento tra tutti i geni espressi in quella cellula, in quanto cellule biologicamente simili tendono a collocarsi vicine nello spazio latente.
Per ciascun gene g nella cellula c, *VeloVI* introduce inoltre una variabile latente discreta k_{cg} che codifica i quattro stati trascrizionali del gene.

- $k_{cg} = \text{on}$: stato di induzione;
- $k_{cg} = \text{ss}^{\text{on}}$: stato stazionario di induzione;
- $k_{cg} = \text{off}$: stato di repressione;
- $k_{cg} = \text{ss}^{\text{off}}$: stato stazionario di repressione.

Per ogni coppia cellula–gene, viene usata la seguente distribuzione di probabilità sui quattro stati:

$$\pi_{cg} \sim \text{Dirichlet}(0.25,0.25,0.25,0.25), \tag{2.27}$$

e quindi campiona lo stato trascrizionale

$$k_{cg} \sim \text{Categorical}(\pi_{cg}). \tag{2.28}$$

Questo significa che, per ciascun gene in ciascuna cellula, il modello non assegna immediatamente uno stato fisso, ma mantiene una descrizione probabilistica dell'appartenenza ai diversi regimi dinamici. I parametri cinetici restano invece stime puntuali per gene, analogamente al modello *scVelo*, ma vengono appresi congiuntamente ai parametri della rete neurale durante il train.

L'architettura di *VeloVI* corrisponde a quella di un VAE ed è costituita da due componenti principali:

- l'**Encoder** è una rete neurale feed-forward completamente connessa che riceve in input le abbondanze spliced e unspliced della cellula c e restituisce la distribuzione a posteriori approssimata di z_c;
- il **Decoder** usa campioni di z_c per produrre, per ciascun gene, il tempo latente, lo stato trascrizionale e le abbondanze attese secondo il modello dinamico.

Nell'encoder, per calcolare la distribuzione a posteriori delle variabili latenti z_c e π_{cg} *VeloVI* utilizza, come anche i metodi precedenti, *variational inference* introducendo una distribuzione approssimante tale che

$$q_\phi(z, \pi \mid x) = \prod_c \Big(q_\phi(z_c \mid x_c) \prod_g q_\phi(\pi_{cg} \mid z_c) \Big), \tag{2.29}$$

parametrizzata da reti neurali con parametri ϕ. Come nei VAE tradizionali, l'encoder restituisce i parametri della distribuzione a posteriori approssimata di z_c,

ossia la media e la varianza della distribuzione gaussiana, nonché i parametri della distribuzione di Dirichlet associata a π_{cg}.

Una volta definito lo stato trascrizionale, il decoder assegna a ciascuna coppia cellula–gene un tempo latente t_{cg}, che però non viene stimato indipendentemente come in *scVelo*. In *VeloVI*, infatti, il tempo dipende dalla rappresentazione latente della cellula z_c tramite reti neurali distinte e indipendenti, una per l'induzione e una per la repressione.

Nel caso di induzione ($k = \text{on}$), si definisce

$$\rho_{cg}^{\text{on}} = [h_{\text{on}}(z_c)]_g, \tag{2.30}$$

dove $h_{\text{on}} : \mathcal{R}^d \to (0,1)^G$ e quindi il tempo latente è parametrizzato come una rete neurale completamente connessa. Si ottiene quindi il tempo come

$$t_{cg}^{\text{on}} = \rho_{cg}^{\text{on}}\, t_g^s, \tag{2.31}$$

dove t_g^s è il tempo di switching del gene g, cioè il tempo in cui il sistema passa dalla fase di induzione alla fase di repressione.

Nel caso di repressione ($k = \text{off}$), si definisce invece

$$\rho_{cg}^{\text{off}} = [h_{\text{off}}(z_c)]_g, \tag{2.32}$$

$$t_{cg}^{\text{off}} = (t_{\max} - t_g^s)\rho_{cg}^{\text{off}} + t_g^s, \tag{2.33}$$

dove $t_{\max}$ è una costante che fissa la scala temporale massima del modello. Analogamente si ha $h_{\text{off}} : \mathcal{R}^d \to (0,1)^G$ anch'essa rete neurale. In questo modo il tempo latente è vincolato a essere coerente con lo stato trascrizionale: se il gene è in induzione, il tempo deve collocarsi prima dello switch; se è in repressione, deve collocarsi dopo.

Per gli stati stazionari $k = \text{ss}^{\text{on}}$ e $k = \text{ss}^{\text{off}}$, il modello considera direttamente il limite delle soluzioni delle equazioni differenziali quando $t \to \infty$, (2.9) e (2.8).

I dati osservati, costituiti dalle abbondanze preprocessate di RNA spliced e unspliced, vengono campionati come distribuzioni normali:

$$u_{cg} \sim \mathcal{N}\Big(\mu_{cg}^{(u)}, (c_{k_{cg}}\sigma_g^{(u)})^2\Big), \tag{2.34}$$

$$s_{cg} \sim \mathcal{N}\Big(\mu_{cg}^{(s)}, (c_{k_{cg}}\sigma_g^{(s)})^2\Big), \tag{2.35}$$

dove le medie $\mu_{cg}^{(u)}$ e $\mu_{cg}^{(s)}$ sono ottenute dalle espressioni teoriche chiuse di RNA unspliced e spliced (2.2) e (2.3), $\sigma_g^{(u)}$ e $\sigma_g^{(s)}$ sono parametri di rumore specifici per

gene e c_k è un fattore di scala dipendente dallo stato. Per tutti gli esperimenti, per lo stato stazionario ss^{off} è stato utilizzato $c_{\text{ss}^{\text{off}}} = 0.1$, mentre per gli altri stati è $c_k = 1$. Questa è una scelta di modellazione euristica adottata per rendere la dispersione minore in un gene che si sta spegnendo.

L'ipotesi di gaussianità è giustificata dal fatto che *VeloVI* non lavora sui conteggi grezzi, come *cell2fate*, ma sui valori preprocessati con la pipeline standard di *scVelo*. La scelta di una likelihood gaussiana riflette quindi l'idea che tali quantità preprocessate si comportino come osservazioni continue rumorose. Poiché lo stato trascrizionale k_{cg} non è osservato, esso viene marginalizzato. Di conseguenza, le distribuzioni condizionate delle abbondanze osservate, dato il vettore latente della cellula $\mathbf{z}_c$ e le probabilità di appartenenza agli stati $\boldsymbol{\pi}_{cg}$, assumono la forma

$$P(u_{cg} \mid z_c, \pi_{cg}) = \sum_{k_{cg}} \pi_{cgk_{cg}} \mathcal{N}\left(\mu_{cg}^{(u)}, (c_{k_{cg}} \sigma_g^{(u)})^2\right), \tag{2.36}$$

$$P(s_{cg} \mid z_c, \pi_{cg}) = \sum_{k_{cg}} \pi_{cgk_{cg}} \mathcal{N}\left(\mu_{cg}^{(s)}, (c_{k_{cg}} \sigma_g^{(s)})^2\right), \tag{2.37}$$

che sono misture di distribuzioni normali.

La funzione obiettivo massimizzata durante il training per stimare i parametri è ELBO regolarizzata, definita come

$$\mathcal{L}_{\text{velo}} = -\text{ELBO} + \lambda \mathcal{L}_{\text{switch}}, \tag{2.38}$$

dove ELBO è la tradizionale Evidence lower Bound definita nei VAE. Il termine aggiuntivo $\mathcal{L}_{\text{switch}}$ ha invece la funzione di regolarizzare la posizione dello switching point nel phase portrait, migliorando la stabilità dell'inferenza

$$\mathcal{L}_{\text{switch}} = \sum_g \left(u_{g0}^{\text{off}} - u_g^*\right)^2 + \left(s_{g0}^{\text{off}} - s_g^*\right)^2$$

dove u_{0g}^{off} e s_{0g}^{off} sono le condizioni iniziali della fase di repressione al tempo di switch. s_g^* e u_g^* sono i valori target ottenuti calcolando la mediana empirica delle quantità selezionando solo le cellule con abbondanze di RNA superiore al 99° percentile. In pratica si ha una penalizzazione aggiuntiva quando il punto di switch previsto dalle equazioni differenziali è lontano da quello suggerito dai dati. Per ottimizzare $\mathcal{L}_{velo}$ è stato usato stochastic gradient descent con Adam optimizer, il parametro λ è stato fissato a 0.2 per tutti gli esperimenti.

Una volta stimato il modello, *VeloVI* consente di calcolare diverse quantità di interesse, che vengono utilizzate sia per la valutazione del modello sia per l'analisi

biologica dei dati. Per ciascuna coppia cellula-gene vengono stimate le probabilità a posteriori di appartenenza ai quattro stati trascrizionali mediante

$$\mathbb{E}_{q_\phi(z_c|u_c,s_c)}\Big[\mathbb{E}_{q_\phi(\pi_{cg}|z_c)}[\pi_{cg}]\Big]. \tag{2.39}$$

Il tempo latente per ciascun gene e ciascuna cellula

$$\mathbb{E}_{q_\phi(z_c|u_c,s_c)}\Big[\mathbb{E}_{q_\phi(\pi_{cg}|z_c)}\Big[t_{cg}^{(k_{cg})}\Big]\Big], \tag{2.40}$$

dove l'attesa esterna integra rispetto alla rappresentazione latente $q_\phi(z_c \mid u_c, s_c)$ ed è stimata con campionamento Monte Carlo, mentre quella interna calcola la media rispetto ai quattro possibili stati trascrizionali k_{cg} ed è calcolata analiticamente. La RNA velocity $v^{(g)}(t^{(k)}, k)$ per un gene g nella cellula c si ottiene come segue:

1. si campiona $z_c \sim q_\phi(z_c \mid u_c, s_c)$;

2. si calcola, per ciascun gene,

$$\mathbb{E}_{q_\phi(\pi_{cg}|z_c)}\Big[v_g(t_{cg}^{(k_{cg})}, k_{cg})\Big]. \tag{2.41}$$

È possibile introdurre un tasso di trascrizione dipendente dal tempo:

$$\alpha^{(k)}(t) = \begin{cases} \alpha_1 - (\alpha_1 - \alpha_0)e^{-\lambda_\alpha t}, & k \in \{\text{on}, ss^{\text{on}}\}, \\ 0, & k \in \{\text{off}, ss^{\text{off}}\}. \end{cases} \tag{2.42}$$

Applicando questa formulazione, si ottengono espressioni chiuse per le abbondanze, che sostituiscono quelle del modello originale e possono essere utilizzate come medie delle distribuzioni gaussiane, permettendo l'estensione della dinamica senza modificare la procedura inferenziale.

La Tabella (2.1) presenta le principali differenze dei metodi elencati precedentemente.

scVelo, data la sua interpretabilità biologica e la relativa semplicità computazionale rimane il framework di riferimento. Tuttavia, l'inferenza gene-wise, l'assenza di una quantificazione dell'incertezza e la forte dipendenza dal preprocessing ne limitano la robustezza.

Caratteristica	**scVelo**	**cell2fate**	**VeloVAE**	**VeloVI**
Modello	ODE gene-wise	Bayesiano modulare	VAE + ODE	VAE prob
Inferenza	EM	SVI	VAE	VAE
Dinamica	On/off discreta	Moduli	α continuo	On/off prob
Tempo latente	Gene-specifico	Globale	Globale	Cellula-gene
Variabili latenti	t_c	t_c	$t_c,\ z_c$	$z_c,\ k_{cg}$
Likelihood	Gaussiana	BN	Gaussiana	Gaussiana
Input	Pre-processing	Conteggi grezzi	Pre-processing	Pre-processing
Incertezza	No	Sì	Sì	Sì
RNA ambientale	No	Sì	Parziale	Parziale
Interpretabilità	Alta	Alta	Intermedia	Intermedia
Costo computazionale	Moderato	Elevato	Elevato	Elevato

Tabella 2.1: Confronto tra i modelli di RNA velocity.

cell2fate introduce una formulazione completamente bayesiana e una decomposizione modulare della trascrizione che consente di descrivere dinamiche multi-step più realistiche. L'utilizzo diretto dei conteggi grezzi e della likelihood binomiale negativa rende inoltre il modello più adatto alla natura discreta e rumorosa dei dati scRNA-seq. Il principale svantaggio è il maggiore costo computazionale.

VeloVAE siccome combina le equazioni cinetiche con i variational autoencoders e le reti neurali risulta essere meno interpretabile biologicamente. Tuttavia, la presenza di uno spazio latente continuo permette di modellare biforcazioni e traiettorie multiple, risultando quindi più flessibile nel rappresentare dinamiche trascrizionali continue.

VeloVI pur mantenendo una struttura on/off del modello dinamico classico permette di propagare l'incertezza fino alla stima della RNA velocity introducendo una formulazione probabilistica grazie ai VAE. L'uso di variabili latenti condivise tra geni migliora la coerenza globale dell'inferenza rispetto ai modelli puramente gene-wise.

Capitolo 3

Dataset

Per testare e validare i vari modelli di RNA velocity le visualizzazioni devono sempre essere contestualizzate con conoscenze biologiche consolidate. Per questo in questa analisi utilizzeremo due dei dataset più utilizzati in letteratura: *pancreatic endocrinogenesis* e *dentate gyrus neurogenesis*, che rappresentano sistemi biologici con dinamiche di differenziamento ben caratterizzate e costituiscono benchmark standard per la valutazione dei metodi di inferenza delle traiettorie cellulari.

Entrambi i dataset sono strutturati secondo il formato standard per dati single-cell *AnnData*, in cui i dati sono organizzati in una matrice principale $X \in \mathbb{R}^{c \times g}$, dove c indica il numero di cellule e g il numero di geni. A questa matrice sono associate diverse annotazioni:

- `obs`: contiene le osservazioni relative alle cellule, come i cluster identificati (`clusters_coarse`, `clusters`);
- `var`: contiene le annotazioni relative ai geni ad esempio quali sono stati selezionati come altamente variabili (`highly_variable_genes`);
- `uns`: raccoglie metadati non strutturati, come colori dei cluster e informazioni già calcolate durante l'analisi (ad esempio PCA);
- `layers`: matrici di RNA spliced e unspliced;
- `obsm`: sono salvate rappresentazioni multivariate delle cellule, come le coordinate nello spazio PCA (`X_pca`) e UMAP (`X_umap`), utili per visualizzazioni e riduzione della dimensionalità;
- `obsp`: matrici che descrivono relazioni tra cellule, come distanze (`distances`) e connettività (`connectivities`) nei grafi.

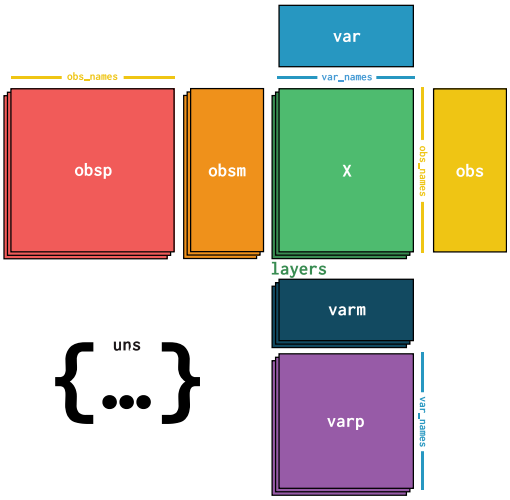


Figura 3.1: Struttura AnnData

Il primo dataset, denominato *pancreatic endocrinogenesis*, deriva dallo studio di Bastidas-Ponce et al. [2019] ed è disponibile nel database Gene Expression Omnibus (GEO) con codice di accesso GSE132188. Si osserva lo sviluppo delle cellule endocrine del pancreas di topo, ottenuto mediante sequenziamento dell'RNA a singola cellula di 36.351 cellule epiteliali embrionali analizzate durante la fase di transizione secondaria (tra i time points E12.5 ed E15.5). È stata utilizzata la piattaforma 10x Genomics Chromium Single Cell 3' Kit V2, una delle tecnologie droplet-based più diffuse. Il dataset distribuito direttamente dalla libreria *scVelo* può essere scaricato direttamente mediante la funzione `scv.datasets.pancreas()`, che restituisce un oggetto `AnnData` contenente 3696 cellule e 27998 geni, le annotazioni cellulari e i conteggi di RNA spliced e unspliced. Tale oggetto costituisce quindi una versione già selezionata e predisposta per l'analisi di RNA velocity.

In questo processo biologico le cellule progenitrici duttali (Ductal), caratterizzate da un comportamento proliferativo e ciclico, danno origine alle cellule progenitrici endocrine (Ngn3 EP, Neurogenin 3 Endocrine Progenitors) che a loro volta si differenziano in quattro principali destini cellulari: cellule α (che producono glucagone), cellule β (produttrici di insulina), cellule δ (responsabili della secrezione di somatostatina) e cellule ε (che secernono grelina). La Figura 3.2 mostra la proiezione bidimensionale UMAP delle cellule, raggruppate nei principali 8 cluster identificati. Siccome la struttura del processo è ben caratterizzata dal punto di vista biologico è possibile ottenere una validazione più solida dei modelli di inferenza delle traiettorie precedentemente descritti. Inoltre, le dimensioni contenute del dataset ottenuto dopo la fase di preprocessing (cellule $\times$ geni $= 3696 \times 2000$) lo rendono computazionalmente gestibile, favorendo il confronto di diversi metodi senza particolari vincoli di scalabilità.

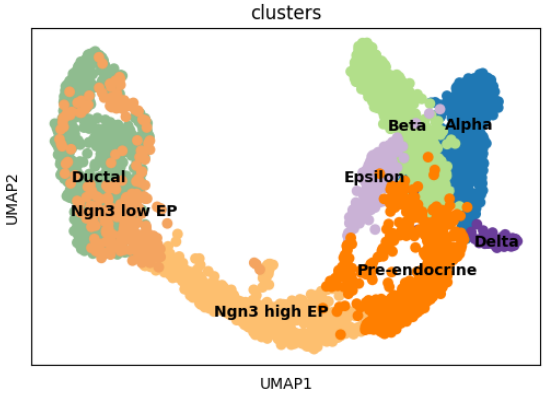


Figura 3.2: Mappa UMAP dello sviluppo epiteliale pancreatico

Il secondo dataset, *dentate gyrus neurogenesis*, descrive lo sviluppo del giro dentato dell'ippocampo nel topo, deriva dallo studio di Hochgerner et al. [2018]. Il dataset originale comprende 5454 cellule raccolte tra i time points P12 e P35 misurati sempre tramite scRNA-seq droplet-based (usando 10x Genomics Chromium Single Cell Kit V1) ed è disponibile nel database GEO nella SuperSeries GSE95753. A differenza del caso precedente, questo sistema biologico è caratterizzato da una struttura di differenziamento più complessa e fortemente ramificata. Il dataset comprende cellule progenitori neurali, neuroni maturi e diversi tipi di cellule gliali, tra cui astrociti e oligodendrocyte precursor cells (OPCs). La caratteristica principale di questo dataset è la presenza di un manifold biologico continuo con multiple biforcazioni, che riflette in modo naturale processi di differenziamento non lineari e multi-lineage complessi. La Figura 3.3 presenta invece la proiezione bidimensionale UMAP dei 14 cluster delle cellule del dataset del giro dentato.

Il principale flusso atteso dal punto di vista biologico corrisponde alla cosiddetta linea granulare in cui le nIPC (neural Intermediate Progenitors Cells) si differenziano in neuroblasti che poi si sviluppano in cellule granulari immature e poi mature. Le cellule radial glia-like, all'origine di questa linea, rappresentano tuttavia un punto di biforcazione: oltre a dare origine alle nIPC e quindi alla linea neurogenica, possono differenziarsi direttamente in astrociti. È inoltre presente una sottolinea ligodendrocitaria (da OPC a OL). Le cellule GABAergiche, pur descritte in letteratura come una popolazione distinta dalla linea granulare principale, non seguono una sequenza di stadi di maturazione interna analoga a quella oligodendrocitaria. Le cellule Cajal-Retzius costituiscono invece una popolazione terminale. Anche questo dataset è stato acquisito tramite la funzione di download integrata nel pacchetto *scVelo*, `scv.datasets.dentategyrus()` e contiene 2930

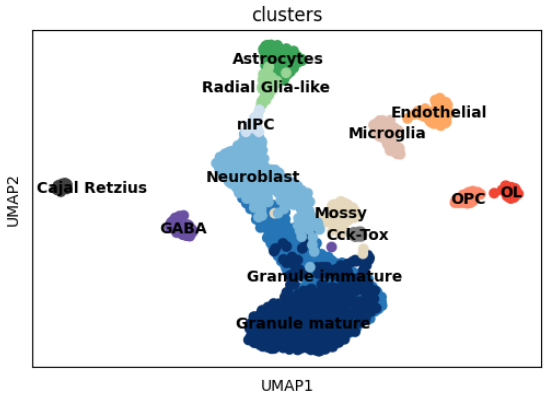


Figura 3.3: Mappa UMAP dello sviluppo del giro dentato dell'ippocampo

cellule e 13913 geni. Dopo il preprocessing e la selezione dei 2000 geni maggiormente variabili, il dataset ha dimensioni pari a 2930 $\times$ 2000.

Nei capitoli successivi verranno applicati i modelli di RNA velocity ai due dataset per confrontarne la capacità di ricostruire le traiettorie di differenziazione note biologicamente. Prima di presentare i risultati, è utile chiarire alcuni aspetti metodologici relativi alla pipeline di analisi dei dati scRNA-seq.

Le tecnologie di scRNA-seq comprendono numerosi protocolli, che differiscono principalmente per il metodo di isolamento delle cellule. Una panoramica delle principali piattaforme è riportata in Lafzi et al. [2018]. Tra questi vi sono approcci plate-based, ad esempio SMART-seq2, che consentono il sequenziamento *full-length* dei trascritti, e approcci droplet-based, come 10x Genomics Chromium, che permettono l'analisi di un numero molto elevato di cellule. Il sequenziamento produce milioni di *reads*, ossia brevi sequenze nucleotidiche. Nella maggior parte dei protocolli, ciascuna read contiene tre informazioni principali: il CB (cell barcode) che identifica la cellula originale, la sequenza di nucleotidi e un UMI (Unique Molecular Identifier) per etichettare univocamente ciascuna molecola. L'UMI viene inserito artificialmente durante la fase di preparazione del campione, questo permette di distinguere le molecole originali dai duplicati generati durante la fase successiva di PCR e di ottenere una analisi più accurata dell'espressione genica a livello cellulare.

Le reads vengono successivamente elaborate mediante pipeline di allineamento e quantificazione. Nel caso dei dati 10x Genomics, una delle pipeline maggiormente utilizzate è *Cell Ranger*, che corregge i barcode cellulari, allinea le reads al genoma di riferimento e le assegna ai rispettivi geni. Gli allineamenti vengono memorizzati in un file in formato BAM (*Binary Alignment Map*), nel quale sono

riportate, per ciascuna read, la posizione genomica e le informazioni associate al barcode cellulare e all'UMI. Le reads vengono assegnate ai rispettivi geni e raggruppate in base alla combinazione cellula-gene-UMI, in un processo noto come UMI collapsing, con l'obiettivo di produrre le due matrici di conteggio. Senza questa procedura, i duplicati introdotti dalla PCR verrebbero erroneamente interpretati come molecole distinte, determinando una sovrastima dell'espressione genica. Il risultato finale è una matrice di conteggio gene $\times$ cellula, in cui ogni elemento rappresenta il numero di molecole di RNA associate a un determinato gene in una specifica cellula. Nel contesto della RNA velocity vengono inoltre generate due matrici distinte contenenti i conteggi delle molecole spliced e unspliced, che costituiscono l'input dei modelli di inferenza della dinamica trascrizionale. La classificazione tra RNA spliced e unspliced rappresenta tuttavia una semplificazione della complessità reale del processo di splicing. Per esempio, consideriamo il trascritto $E_1I_1E_2I_2E_3$, dove E_i rappresenta un esone e I_i un introne. Lo splicing, generalmente, può attraversare stati intermedi, come $E_1E_2I_2E_3$, prima di raggiungere la forma completamente matura $E_1E_2E_3$. Bisogna quindi evidenziare che già in questa fase di raccolta dati, la classificazione di questi stati intermedi in una delle due categorie rende il modello incompleto, in quanto la cinetica dello splicing non sempre è riducibile a un solo passaggio.

Oltre alla pipeline costituita da *Cell Ranger* e *velocyto*, sono disponibili diversi strumenti per la quantificazione delle molecole spliced e unspliced. Tra questi strumenti ci sono STARsolo, basato sull'algoritmo STAR Dobin et al. [2013] che esegue un allineamento completo delle reads, kallisto|bustools che implementa uno pseudo allineamento, Bray et al. [2016], Melsted et al. [2019]. In Du et al. [2020] è presente un confronto sistematico tra i due: mostrano come STARsolo rilevi generalmente un numero maggiore di geni e raggiunga tassi di allineamento più elevati, ma con un consumo di memoria e di tempi di esecuzione superiori rispetto a kallisto|bustools. Un'ulteriore pipeline per protocolli droplet-based è dropEst Petukhov et al. [2018], progettato per migliorare l'accuratezza della stima dei conteggi molecolari attraverso la correzione degli errori introdotti durante il sequenziamento. Un confronto sistematico tra alcune pipeline di quantificazione è stato condotto da Soneson et al. [2021]. Gli autori mostrano che le differenze introdotte dai metodi di quantificazione non sono trascurabili e possono propagarsi fino alla stima della RNA velocity: in particolare si osserva il fenomeno del **backflow**. In alcuni dataset alcune pipeline producono velocità orientate dalle cellule mature verso le cellule progenitrici, cioè nella direzione opposa rispetto al differenziamento noto biologicamente. Questo risultato mostra come la quantificazione delle abbondanze possa influenzare significativamente le stime di RNA velocity e l'interpretazione biologica dei risultati.
Per entrambi i dataset, i dati prodotti mediante tecnologia 10x Genomics sono

stati inizialmente elaborati con Cell Ranger, che ha generato le matrici di conteggi UMI e i file BAM. I file BAM sono stati successivamente processati mediante *velocyto* per ottenere le matrici di conteggi spliced e unspliced.

Dopo la quantificazione delle matrici spliced e unspliced e la stima della RNA velocity nello spazio ad alta dimensionalità dei geni, i vettori vengono generalmente proiettati in una rappresentazione bidimensionale ottenuta mediante algoritmi di riduzione della dimensionalità, come UMAP o t-SNE. Anche se il seguente aspetto è già stato discusso precedentemente, è utile ribadire le criticità che emergono durante questa fase. Tale passaggio è indispensabile per la visualizzazione e l'interpretazione dei risultati, ma introduce anche diverse fonti di instabilità. Innanzitutto, non viene preservata esattamente la struttura geometrica e le distanze originali dello spazio dei geni, con la conseguenza che le relazioni tra le cellule possano essere distorte o enfatizzate. In secondo luogo, i risultati dipendono fortemente dalla scelta dell'algoritmo. Come vedremo questo rende le inferenze potenzialmente non riproducibili o sensibili a scelte tecniche come quella degli iperparametri. Bisogna precisare che la visualizzazione dei risultati verrà effettuata su UMAP, che è non lineare, pertanto potrebbe alterare la direzione e l'intensità dei vettori stessi.

Capitolo 4

Analisi comparativa

In questo capitolo vengono presentati e confrontati i risultati ottenuti applicando i metodi *scVelo*, *cell2fate*, *VeloVI* e *VeloVAE* (nelle sue varianti standard, *FullVB* e *BrODE*) ai dataset di endocrinogenesi pancreatica e di neurogenesi del giro dentato descritti nel Capitolo 3. Inizialmente si presentano alcune metriche di valutazione che permettono di misurare in modo oggettivo la correttezza delle transizioni inferite e la coerenza locale delle velocity stimate, in accordo con recenti studi di benchmarking dei metodi di RNA velocity Luo et al. [2025] e Huang et al. [2026]. Successivamente i risultati di queste metriche vengono discussi congiuntamente all'osservazione diretta dei campi di velocità proiettati nello spazio UMAP e dell'ordinamento temporale delle cellule fornito dal tempo latente. In questa fase, particolare attenzione è dedicata alla direzione dei flussi cellulari ricostruiti e all'eventuale presenza di inversioni nella dinamica ricostruita. L'obiettivo è valutare non solo la qualità delle stime prodotte da ciascun metodo, ma anche la loro stabilità rispetto a scelte di inizializzazione e di iperparametri, aspetto centrale nella discussione avviata nel Capitolo 1.

4.1 Background matematico delle metriche di valutazione

Prima di elencare le metriche di valutazione è utile introdurre la similarità coseno.

Dati due vettori non nulli $\mathbf{a}, \mathbf{b} \in \mathbb{R}^q$, la similarità coseno è definita come

$$\cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^\top \mathbf{b}}{\|\mathbf{a}\|_2 \|\mathbf{b}\|_2} \tag{4.1}$$

e assume valori nell'intervallo $[-1,1]$. Misura l'allineamento angolare tra i due vettori, indipendentemente dalla loro lunghezza. Un valore prossimo a 1 indica che

i due vettori sono orientati nella stessa direzione, prossimo a 0 che sono approssimativamente ortogonali, mentre prossimo a -1 che sono orientati in direzioni opposte.

Poiché la norma compare al denominatore, la similarità coseno non è definita quando almeno uno dei due vettori è nullo. Per questo motivo, nell'implementazione vengono escluse le cellule con velocity nulla e le direzioni di spostamento aventi norma nulla.

Cross-Boundary Direction Correctness

La *Cross-Boundary Direction Correctness*, indicata con CBDir, misura se le velocity delle cellule poste al confine tra due cluster siano orientate nella direzione della transizione biologica attesa.

Per ogni cellula i, si definisce l'insieme relativo di vicini come

$$\mathcal{N}(i) = \{j \in \mathcal{C} : (i, j) \in \mathcal{E}\},$$

dove $\mathcal{C}$ è l'insieme delle cellule e $\mathcal{E}$ l'insieme degli archi del grafo k-NN delle cellule, costruito nello spazio dei vicini più prossimi già utilizzato nella fase di preprocessing (`adata.obsp['connectivities']`).
Sia inoltre $\mathcal{K}$ l'insieme dei cluster. Si consideri una transizione $k_s \longrightarrow k_t$, dove $k_s \in \mathcal{K}$ è il cluster sorgente e $k_t \in \mathcal{K}$ è il cluster destinazione.
Per una cellula i appartenente al cluster sorgente, cioè tale che $k_i = k_s$, si definisce l'insieme dei vicini appartenenti al cluster destinazione come

$$\mathcal{N}_{k_t}(i) = \{j \in \mathcal{N}(i) : k_j = k_t\}. \tag{4.2}$$

Le cellule del cluster sorgente che possiedono almeno un vicino nel cluster destinazione costituiscono l'insieme delle cellule di confine:

$$\mathcal{B}_{k_s \to k_t} = \{i \in \mathcal{C} : k_i = k_s,\, \mathcal{N}_{k_t}(i) \neq \varnothing\}. \tag{4.3}$$

Sia $\mathbf{x}_i \in \mathbb{R}^2$ la posizione della cellula i nello spazio di embedding, e sia $\widetilde{\mathbf{v}}_i$ la velocity della cellula i proiettata nello stesso spazio. Per ogni cellula $i \in \mathcal{B}_{k_s \to k_t}$ e per ogni vicino $j \in \mathcal{N}_{k_t}(i)$, si considera il vettore

$$\mathbf{d}_{ij} = \mathbf{x}_j - \mathbf{x}_i, \tag{4.4}$$

che rappresenta la direzione dalla cellula sorgente i verso una cellula del cluster destinazione. Il punteggio della singola cellula è dato dalla media delle similarità coseno tra la velocity proiettata $\widetilde{\mathbf{v}}_i$ e le direzioni verso i vicini del cluster destinazione:

$$\mathrm{CBDir}_i\,(k_s \to k_t) = \frac{1}{|\mathcal{N}_{k_t}(i)|} \sum_{j \in \mathcal{N}_{k_t}(i)} \frac{\widetilde{\mathbf{v}}_i^{\top} \mathbf{d}_{ij}}{\|\widetilde{\mathbf{v}}_i\|_2 \|\mathbf{d}_{ij}\|_2}. \tag{4.5}$$

La CBDir totale associata alla transizione $k_s \to k_t$ è ottenuta mediando i punteggi delle cellule di confine

$$\mathrm{CBDir}(k_s \to k_t) = \frac{1}{|\mathcal{B}_{k_s \to k_t}|} \sum_{i \in \mathcal{B}_{k_s \to k_t}} \mathrm{CBDir}_i\,(k_s \to k_t). \tag{4.6}$$

Se tale valore è prossimo a 1 le velocity sono orientate dal cluster sorgente verso il cluster destinazione, in accordo con la transizione biologica attesa; se è 0 non c'è una direzione prevalente, mentre se il valore è negativo le velocity tendono a puntare nella direzione opposta rispetto alla transizione attesa.

Sia

$$\mathcal{T} = \left\{(k_s^{(1)}, k_t^{(1)}), \ldots, (k_s^{(M)}, k_t^{(M)})\right\}$$

l'insieme delle M transizioni biologiche considerate. Indicando con $\mathcal{T}^* \subseteq \mathcal{T}$ le transizioni per cui il punteggio è definito, il valore complessivo è

$$\mathrm{CBDir}_{\mathrm{overall}} = \frac{1}{|\mathcal{T}^*|} \sum_{(k_s,k_t) \in \mathcal{T}^*} \mathrm{CBDir}(k_s \to k_t). \tag{4.7}$$

La media è calcolata rispetto alle transizioni e non rispetto alle singole cellule. Di conseguenza, ogni transizione contribuisce allo stesso modo al risultato complessivo, indipendentemente dal numero di cellule presenti sul confine.

In-Cluster Coherence

L'*In-Cluster Coherence*, indicata con ICCoh, misura la coerenza locale delle velocity tra cellule vicine appartenenti allo stesso cluster.

Per una cellula i, si definisce l'insieme dei vicini appartenenti allo stesso cluster come

$$\mathcal{N}_{k_i}(i) = \{j \in \mathcal{N}(i) : k_j = k_i\}\,. \tag{4.8}$$

Il punteggio della cellula i è definito come la similarità coseno media tra la sua velocity e quelle dei vicini appartenenti allo stesso cluster

$$\mathrm{ICCoh}_i = \frac{1}{|\mathcal{N}_{k_i}(i)|} \sum_{j \in \mathcal{N}_{k_i}(i)} \frac{\mathbf{v}_i^\top \mathbf{v}_j}{\|\mathbf{v}_i\|_2 \|\mathbf{v}_j\|_2}. \tag{4.9}$$

Per ogni cluster $k \in \mathcal{K}$, si introduce l'insieme delle cellule valide

$$\mathcal{I}_k = \{i \in \mathcal{C} : k_i = k,\, \mathcal{N}_k(i) \neq \varnothing,\, \|\mathbf{v}_i\|_2 > 0\}\,. \tag{4.10}$$

La coerenza relativa al cluster k è quindi

$$\mathrm{ICCoh}(k) = \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \mathrm{ICCoh}_i \,. \tag{4.11}$$

Valori elevati indicano che cellule vicine appartenenti allo stesso cluster presentano velocity orientate in maniera concorde. Mentre valori negativi indicano che le velocity delle cellule vicine tendono ad avere direzioni opposte.

Indicando con

$$\mathcal{K}^* = \{k \in \mathcal{K} : \mathrm{ICCoh}(k) \text{ è definita}\}$$

l'insieme dei cluster con punteggio valido, la coerenza complessiva è

$$\mathrm{ICCoh}_{\mathrm{overall}} = \frac{1}{|\mathcal{K}^*|} \sum_{k \in \mathcal{K}^*} \mathrm{ICCoh}(k). \tag{4.12}$$

Anche in questo caso si tratta di una media non pesata. Ogni cluster contribuisce quindi allo stesso modo al punteggio complessivo, indipendentemente dal numero di cellule che contiene.

Consistency score basato sul grafo condiviso

Il consistency score, CS, basato sul grafo condiviso misura la coerenza locale delle velocity tra una cellula e tutte le cellule appartenenti al suo vicinato, senza imporre che queste appartengano allo stesso cluster.

Per ogni cellula i, si definisce

$$\mathrm{CS}_i = \frac{1}{|\mathcal{N}^*(i)|} \sum_{j \in \mathcal{N}^*(i)} \frac{\mathbf{v}_i^\top \mathbf{v}_j}{\|\mathbf{v}_i\|_2 \|\mathbf{v}_j\|_2}, \tag{4.13}$$

dove

$$\mathcal{N}^*(i) = \{j \in \mathcal{N}(i) : \|\mathbf{v}_j\|_2 > 0\} \tag{4.14}$$

è l'insieme dei vicini con velocity non nulla.
Si definisce inoltre l'insieme delle cellule valide

$$\mathcal{I}^* = \{i \in \mathcal{C} : \|\mathbf{v}_i\|_2 > 0, \mathcal{N}^*(i) \neq \varnothing\} \,. \tag{4.15}$$

Il consistency score complessivo è

$$\mathrm{CS}_{\mathrm{graph}} = \frac{1}{|\mathcal{I}^*|} \sum_{i \in \mathcal{I}^*} \mathrm{CS}_i \,. \tag{4.16}$$

Questa metrica è strettamente collegata alla ICCoh. La differenza principale riguarda la scelta dei vicini. La ICCoh misura infatti la regolarità locale del campo all'interno dei cluster, mentre il consistency score basato sul grafo può considerare anche le relazioni tra cellule appartenenti a cluster differenti.

Consistency score con correlazione di Pearson

Il consistency score basato su correlazione di Pearson, CS_{Pearson}, è una variante di CS_{graph} in cui, al posto della similarità coseno pura, si utilizza la correlazione di Pearson tra la velocity di ogni cellula e quella dei suoi vicini. Si osservi che ciò equivale a centrare ciascun vettore sottraendo la propria media prima di calcolarne la similarità coseno.

Sia $\bar{v}_i$ la media dei valori di $\mathbf{v}_i$ sui geni, e sia

$$\mathbf{v}_i^c = \mathbf{v}_i - \bar{v}_i \mathbf{1} \tag{4.17}$$

la velocity centrata della cellula i, dove $\mathbf{1}$ è il vettore di uno di dimensione pari al numero di geni. Il punteggio della cellula i è definito come

$$\mathrm{CS}_i^{\text{Pearson}} = \frac{1}{|\mathcal{N}^*(i)|} \sum_{j \in \mathcal{N}^*(i)} \frac{(\mathbf{v}_i^c)^\top \mathbf{v}_j^c}{\|\mathbf{v}_i^c\|_2 \|\mathbf{v}_j^c\|_2}, \tag{4.18}$$

dove $\mathcal{N}^*(i)$ è l'insieme dei vicini di i con velocity non nulla, definito come in (4.14). Il punteggio complessivo è

$$\mathrm{CS}_{\text{Pearson}} = \frac{1}{|\mathcal{I}^*|} \sum_{i \in \mathcal{I}^*} \mathrm{CS}_i^{\text{Pearson}}, \tag{4.19}$$

con $\mathcal{I}^*$ definito come in (4.15).

A differenza di CS_{graph}, il centraggio in (4.17) rimuove dal vettore la componente lungo la direzione uniforme dello spazio dei geni (corrispondente al livello medio di espressione della velocity). Il confronto è determinato dalle variazioni relative delle velocity tra i geni, anziché anche da eventuali componenti globali condivise da tutti i geni. Nella documentazione ufficiale di *scVelo* la coerenza è descritta in termini di correlazione tra vettori di velocity vicini, implementata nella funzione `scv.tl.velocity_confidence`. La differenza rispetto al consistency score calcolato da *scVelo* risiede nel grafo di vicinato impiegato ($\mathcal{N}^*(i)$, basato su `adata.obsp['connectivities']`, anziché sul grafo delle distanze usato internamente da scVelo) e nell'assenza di filtri di qualità sui geni, garantendo così coerenza metodologica con CS_{graph}, CBDir e ICCoh.

Correlazione tra tempo latente e ordine biologico noto

Questa metrica valuta se il tempo latente stimato da un metodo sia concorde con l'ordine di sviluppo biologico noto tra i cluster.

Sia $\mathcal{O} = (k_1, k_2, \ldots, k_M)$ la sequenza ordinata dei M cluster secondo la traiettoria biologica attesa, dal più precoce al più tardivo, e sia

$$r(k) = m \quad \text{se } k = k_m \in \mathcal{O} \tag{4.20}$$

il rango del cluster k nell'ordine $\mathcal{O}$. Per ogni cellula i il cui cluster k_i compare in $\mathcal{O}$, sia t_i il tempo latente stimato dal metodo. L'insieme delle cellule valide è

$$\mathcal{C}^* = \{i \in \mathcal{C} : k_i \in \mathcal{O},\ t_i \text{ definito}\}\,. \tag{4.21}$$

Si definiscono pertanto i vettori

$$\mathbf{r} = (r(k_i))_{i \in \mathcal{C}^*}\,, \qquad \mathbf{t} = (t_i)_{i \in \mathcal{C}^*}\,,$$

entrambi di dimensione pari al numero di cellule considerate. Il vettore $\mathbf{r}$ contiene, per ciascuna cellula, il rango del cluster di appartenenza secondo l'ordine biologico atteso; cellule appartenenti allo stesso cluster condividono lo stesso valore. Il vettore $\mathbf{t}$ contiene i tempi latenti stimati dal metodo.

La concordanza tra tempo latente e ordine biologico è misurata tramite il coefficiente di correlazione di **Spearman** per i ranghi

$$\rho = \text{Spearman}(\mathbf{r}, \mathbf{t}) \tag{4.22}$$

dove $\text{Spearman}(\cdot, \cdot)$ indica il coefficiente di correlazione di Spearman tra le due serie di valori. Un valore di ρ prossimo a 1 indica che il tempo latente ordina correttamente le cellule lungo la traiettoria biologica attesa; mentre un valore prossimo a -1 indica un ordinamento invertito.

Quando la traiettoria biologica presenta biforcazioni (ossia un cluster da cui originano più linee distinte, come nel caso del dataset pancreas), la correlazione ρ viene calcolata separatamente per ciascuna sotto-linea, ciascuna rappresentata da una propria sequenza ordinata $\mathcal{O}$, e il punteggio complessivo è ottenuto mediando i valori di ρ validi tra tutte le sotto-linee considerate.

Si noti che questo approccio si ispira alla pratica adottata in letteratura per la valutazione dell'accuratezza del tempo latente, in cui la correlazione di Spearman viene calcolata tra il tempo inferito da un metodo e un riferimento noto a priori – ad esempio il tempo di cattura sperimentale, quando le cellule sono state campionate a timepoint distinti, oppure un punteggio di potenziale differenziativo calcolato in modo indipendente dai dati stessi. In assenza di timepoint sperimentali reali, come nel caso dei dataset utilizzati in questo lavoro, si è scelto di sostituire tale riferimento con il rango del cluster nell'ordine biologico noto, unica informazione a priori disponibile. Questa scelta implica una spaziatura uniforme tra stati successivi; tale assunzione non compromette la validità della correlazione di Spearman, poiché dipende unicamente dalla correttezza dell'ordinamento e non dalla distanza numerica tra i ranghi assegnati.

Transizioni biologiche di riferimento

Le metriche CBDir e la correlazione tra tempo latente e ordine biologico richiedono entrambe la definizione a priori di un insieme di transizioni attese tra cluster, sulla base delle conoscenze biologiche consolidate. Le transizioni utilizzate in questo lavoro sono state definite sulla base della letteratura di riferimento per ciascun dataset.

Per il dataset *pancreatic endocrinogenesis*, la traiettoria di riferimento segue lo schema descritto da Bastidas-Ponce et al. [2019]. Le transizioni considerate sono le seguenti

$$\text{Ductal} \rightarrow \text{Ngn3 low EP} \rightarrow \text{Ngn3 high EP} \rightarrow \text{Pre-endocrine} \rightarrow \{\alpha, \beta, \delta, \epsilon\}.$$

Per il dataset *dentate gyrus neurogenesis*, la linea principale segue lo schema riportato da Hochgerner et al. [2018]:

$$\text{nIPC} \rightarrow \text{Neuroblast} \rightarrow \text{Granule immature} \rightarrow \text{Granule mature}.$$

A questa si affiancano la sotto-linea degli astrociti Radial Glia-like → Astrocytes e quella oligodendrocitaria OPC → OL, anch'essa parte del sistema ma indipendente dalla linea granulare.

Le popolazioni GABAergiche, Cck-Tox, Mossy, Cajal-Retzius, Microglia ed Endothelial non sono state incluse tra le transizioni valutate, in quanto non appartengono alla linea di differenziamento principale o non presentano una sequenza di maturazione interna analoga a quella granulare o oligodendrocitaria, come discusso nel Capitolo 3.

4.2 Configurazioni sperimentali

Per ciascun metodo sono state adottate le configurazioni sperimentali riassunte in Tabella 4.1, descritte nel dettaglio di seguito.

Per gli esperimenti con *scVelo* è stato utilizzato il modello dinamico. Il preprocessing è stato effettuato selezionando i geni con almeno 20 conteggi condivisi e mantenendo i 2000 geni maggiormente variabili. I momenti delle abbondanze spliced e unspliced sono stati calcolati utilizzando 30 componenti principali e un grafo costituito dai 30 vicini più prossimi.

Per *cell2fate* sono stati mantenuti i geni con almeno 20 conteggi condivisi tra le componenti spliced e unspliced e sono stati selezionati i 2000 geni maggiormente variabili. L'addestramento è stato eseguito utilizzando un massimo di 500 epoche, con un mini-batch di 1000, learning rate pari a 0.01 e ottimizzatore `ClippedAdam`. I restanti parametri sono stati mantenuti ai valori di default dell'implementazione.

Per *VeloVI* è stato adottato lo stesso preprocessing utilizzato per *scVelo*. È stata inoltre applicata la funzione `preprocess_data` fornita dal pacchetto *VeloVI*,

Metodo	**Versione**	**Preprocessing**	**Parametri**	**Device**
scVelo	0.2.4	20 conteggi; 2000 HVG; normalizzazione, log e momenti; $n_{\text{PC}} = 30$; $n_{\text{neighbors}} = 30$.	Parametri di inferenza di default.	CPU
cell2fate	0.1a0	20 conteggi; 2000 HVG; conteggi grezzi.	epoche = 500; batch = 1000; lr = 0.01.	CPU
VeloVI	0.4.0	Come *scVelo*; `preprocess_data` (min-max + filtro R^2).	$d_z = 10$; epoche = 500; batch = 256; lr = 0.01; weight decay = 0.01.	CPU
VeloVAE	0.1.2	Come *scVelo*.	$d_z = 5$; $T_{\max} = 20$; epoche = 1000; post-training = 20×500; lr = 10^{-4p-3}; $\text{lr}_{\text{ODE}} = 8\,\text{lr}$.	CPU
FullVB	0.1.2	Come *VeloVAE* standard.	Come *VeloVAE* standard; prior log-normali su $\alpha_g, \beta_g, \gamma_g$: media 0, dev. std. 1 ($\alpha_g$), 0.5 ($\beta_g$, γ_g).	CPU
BrODE	0.1.2	Come *VeloVAE* standard.	Inizializzato da cluster, tempo latente e stato cellulare di *VeloVAE* standard.	CPU

Tabella 4.1: Configurazioni sperimentali utilizzate per l'applicazione dei metodi.

secondo la procedura standard indicata dagli autori: le abbondanze spliced `Ms` e unspliced `Mu` vengono scalate min-max indipendentemente, e vengono mantenuti per l'inferenza solo i geni con coefficiente R^2 e tasso di degradazione γ positivi, stimati tramite il modello deterministico di *scVelo*, oltre a soddisfare i criteri di qualità del fit codificati nella colonna `velocity_genes` di *scVelo*. È stata adottata l'architettura predefinita, nella quale lo stato cellulare è rappresentato in uno spazio latente di dimensione 10. *VeloVI* è stato addestrato per un massimo di 500 epoche con un learning rate pari a 0.01, un coefficiente di weight decay di 0.01 e mini-batch di 256 cellule, con il criterio di *early stopping* attivo.

Per l'inferenza della RNA velocity con *VeloVAE* è stata utilizzata la versione continua standard, nella quale i parametri cinetici vengono stimati come valori puntuali. Il modello è stato inizializzato fissando il tempo massimo della dinamica a $T_{max} = 20$. La dimensione dello spazio latente utilizzato per rappresentare lo stato trascrizionale delle cellule è stata fissata a $d_z = 5$, valore suggerito dagli autori per dataset caratterizzati da un numero limitato di popolazioni cellulari. La prima fase prevede un massimo di 1000 epoche, mentre la successiva fase di raffinamento può comprendere fino a 20 iterazioni, ciascuna costituita da un massimo di 500 epoche. In entrambe le fasi il numero effettivo di epoche può essere inferiore, poiché l'ottimizzazione è regolata da criteri di *early stopping*. Nella configurazione predefinita di *VeloVAE*, i learning rate non sono fissati a valori costanti, ma vengono determinati in funzione della sparsità p del dataset. Per la variante continua utilizzata in questa analisi, il learning rate generale è posto pari a $\eta = 10^{-4p-3}$. Lo stesso valore viene impiegato nella fase di post-training, mentre per l'ottimizzazione dei parametri associati alla componente ODE viene utilizzato un learning rate pari a $\eta_{\text{ODE}} = 8\eta$.

Sono state considerate anche due estensioni proposte nello stesso framework di *VeloVAE*: il modello *FullVB* e il modello *BrODE*. *FullVB* estende l'inferenza variazionale anche ai tassi di trascrizione, splicing e degradazione. Per ciascun gene vengono pertanto approssimate le distribuzioni posteriori di α_g, β_g e γ_g, utilizzando distribuzioni log-normali come prior. Sono stati usati i valori di default: per α_g sono state utilizzate media 0 e deviazione standard 1, mentre per β_g e γ_g sono state utilizzate media 0 e deviazione standard 0.5.

Il modello *BrODE* adotta invece una struttura ODE ramificata, nella quale i parametri cinetici possono variare tra i diversi tipi cellulari. Tale modello è stato inizializzato utilizzando le annotazioni dei cluster, il tempo latente e lo stato cellulare precedentemente inferiti mediante *VeloVAE* standard. I suoi risultati non rappresentano quindi una stima completamente indipendente, ma un raffinamento strutturato delle dinamiche ricostruite dal modello principale.

Per garantire la riproducibilità degli esperimenti, il seed globale è stato fissato a 0 per *scVelo*, *VeloVI* e *cell2fate*. Per *VeloVAE* e le sue varianti, invece, il seed è 2022, in accordo con la configurazione proposta dagli autori.

4.3 Risultati

Sulla base delle metriche introdotte, si procede al confronto tra i metodi, affiancando ai risultati quantitativi l'osservazione diretta dei campi di velocità nello spazio UMAP e del tempo latente stimato. Per una rappresentazione omogenea dei risultati, i campi di velocità ottenuti con i diversi approcci sono stati visualizzati utilizzando la funzione `scv.pl.velocity_embedding_stream` della libreria *scVelo*, applicata ai rispettivi vettori di velocità stimati da ciascun metodo. Il confronto è finalizzato a evidenziare analogie e differenze nella direzione e nella coerenza biologica dei flussi cellulari ricostruiti. Particolare attenzione viene posta alla presenza di eventuali fenomeni di **backflow**, nei quali le velocità risultano orientate dalle popolazioni cellulari mature verso le rispettive popolazioni progenitrici, in direzione opposta rispetto al processo di differenziamento atteso.

4.3.1 Endocrinogenesi pancreatica

Dalla Tabella 4.2 emerge che nessun metodo prevale significativamente sugli altri per tutte le metriche considerate. Invece nella Figura 4.1 per *cell2fate*, *VeloVAE* (in tutte e tre le versioni), si osserva un fenomeno di backflow particolarmente marcato rispetto alla traiettoria di differenziamento nota in letteratura per il pancreas (secondo la quale le cellule progenitrici Ductal e Ngn3 EP danno origine, attraverso lo stadio Pre-endocrine, alle quattro popolazioni endocrine mature α, β, δ, ε).

Metodo	**CBDir**	**ICCoh**	**CS graph**	**CS Pearson**	**Latent time corr.**
scVelo	0.476	0.798	0.798	0.796	0.935
cell2fate	-0.098	0.950	0.939	0.941	0.871
VeloVI	0.518	0.959	0.957	0.956	0.866
VeloVAE	0.114	0.710	0.722	0.722	-0.580
FullVB	0.126	0.888	0.871	0.871	0.553
BrODE	0.224	0.934	0.806	0.806	-0.580

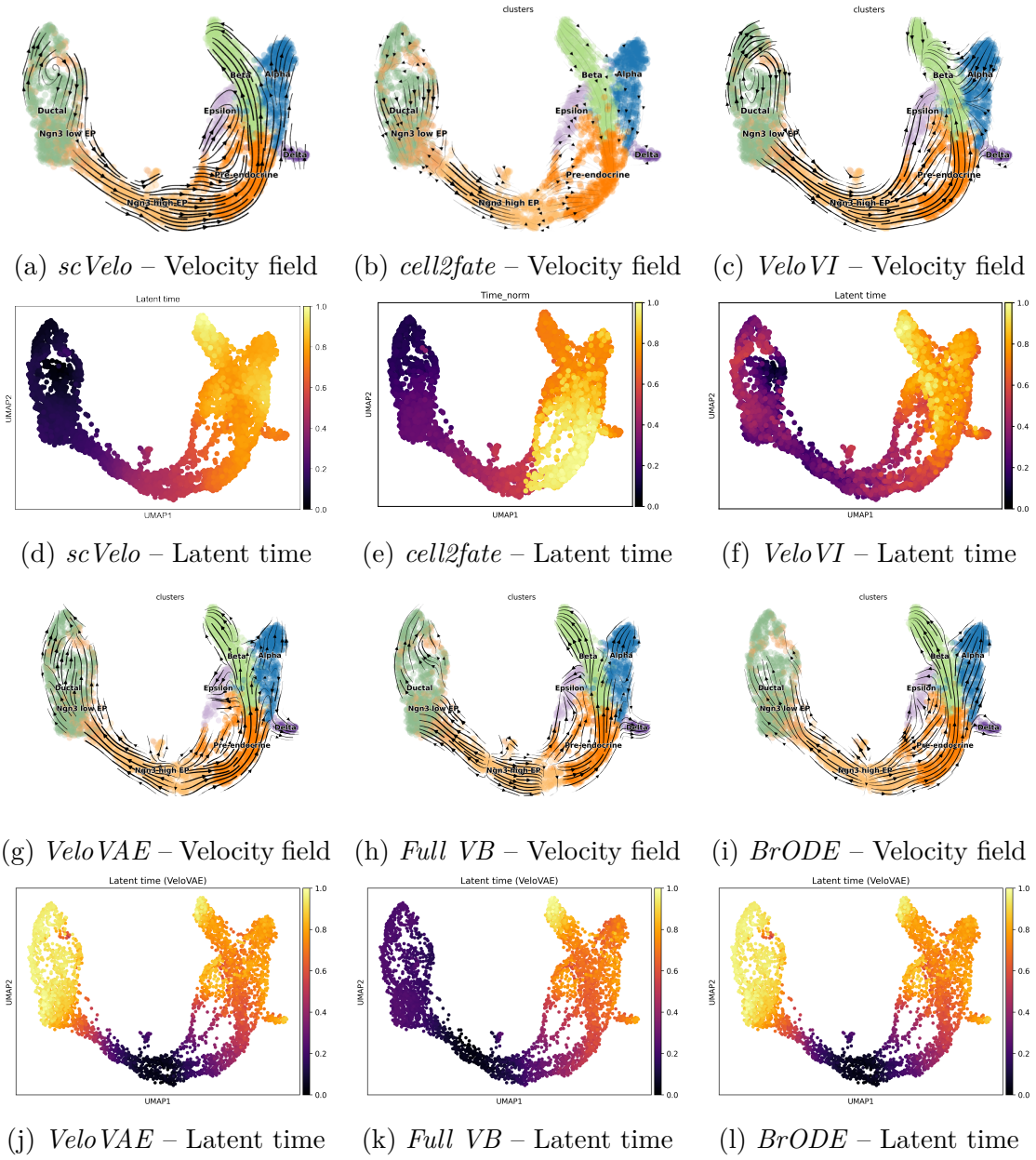
Tabella 4.2: Risultati delle metriche di valutazione per i metodi di RNA velocity applicati al dataset del pancreas.

VeloVI ottiene la migliore CBDir (0.518) ma una correlazione temporale modesta (0.866); coerentemente, il campo di velocità mostra solo un lieve backflow, confinato alla parte superiore dei cluster β e α Figura 4.1c, che non altera la direzione predominante del flusso. *VeloVI* infatti introducendo uno stato cellulare latente integra l'informazione proveniente da più geni e descrive per ciascuna coppia cellula-gene la dinamica trascrizionale mediante una distribuzione a mistura sui quattro possibili stati.

scVelo presenta la correlazione temporale più elevata (0.935), coerente con un campo di velocità visivamente ordinato, sebbene la CBDir (0.476) sia leggermente inferiore a quella di *VeloVI*.

cell2fate mostra una maggiore instabilità: innanzitutto la CBDir è negativa (-0.098) mentre i valori di ICCoh e CS sono alti. Dall'analisi del campo di velocità si vede come le Ngn3 EP procedono prevalentemente verso le Pre-endocrine, in accordo con la traiettoria attesa, ma con qualche fenomeno di backflow locale (in particolare tra le Ngn3 low e Ngn3 high). A partire dalle cellule Pre-endocrine, però, il campo di velocità mostra diverse discontinuità: da un lato questa regione presenta una carenza di frecce, dall'altro le popolazioni endocrine mature (α, β) presentano flussi diretti verso le Pre-endocrine anziché in direzione opposta. Questa inversione è confermata anche dalla corrispondente Figura 4.2d del tempo latente: alla popolazione Pre-endocrine vengono assegnati i valori di tempo latente più elevati tra tutti i cluster, collocandola come l'ultima a differenziarsi lungo la traiettoria.

VeloVAE ottiene una CBDir bassa (0.114) e una correlazione temporale negativa (-0.580) e mostra un campo di velocità con backflow in più punti della traiettoria. In Figura 4.1g dalla popolazione Ngn3 high EP si osservano due flussi: uno diretto all'indietro, verso Ngn3 low EP e, più a monte, verso Ductal e uno invece correttamente orientato verso Pre-endocrine e, successivamente, verso le popolazioni cellulari mature. Un secondo fenomeno di backflow si osserva localmente



(a) *scVelo* – Velocity field (b) *cell2fate* – Velocity field (c) *VeloVI* – Velocity field

(d) *scVelo* – Latent time (e) *cell2fate* – Latent time (f) *VeloVI* – Latent time

(g) *VeloVAE* – Velocity field (h) *Full VB* – Velocity field (i) *BrODE* – Velocity field

(j) *VeloVAE* – Latent time (k) *Full VB* – Latent time (l) *BrODE* – Latent time

Figura 4.1: Confronto delle stime di RNA velocity e del tempo latente ottenute con *scVelo*, *cell2fate*, *VeloVI* e *VeloVAE* (standard, *FullVB*, *BrODE*) sul dataset del pancreas.

tra le popolazioni mature, con un flusso diretto da β verso ϵ, anch'esso in direzione opposta. Infine, all'interno delle α i vettori di velocità risultano poco distinguibili tra loro, con direzioni non chiaramente identificabili. L'analisi del tempo latente

conferma ulteriormente l'incoerenza del modello su questo dataset: alla popolazione Ductal vengono assegnati i valori di tempo latente più elevati tra tutti i cluster, collocandola erroneamente come l'ultima in questa scala di tempo; di conseguenza la sua correlazione temporale risulta negativa (-0.580).

Le altre due formulazioni di *VeloVAE* ricostruiscono una progressione temporale complessivamente simile a quella standard. Il modello *BrODE*, essendo inizializzato a partire dal tempo e dallo stato latente già inferiti da *VeloVAE*, presenta una struttura del campo di velocità sostanzialmente simile a quella del modello standard, ereditandone i medesimi backflow. Il modello *Full VB* mostra anch'esso la persistenza dello stesso campo di velocità osservato nella formulazione standard. Tuttavia, l'analisi del tempo latente rivela che le Ductal si differenziano dopo le Ngn3 high EP, ma molto prima delle β e questo si riflette in una correlazione del tempo più alta rispetto agli altri due (0.553).

La Tabella 4.3 scompone il valore aggregato di CBDir nelle singole transizioni biologiche. La transizione Ngn3 low EP $\rightarrow$ Ngn3 high EP evidenzia una netta separazione tra due gruppi di metodi. *scVelo* e *VeloVI* ottengono valori positivi (0.881 e 0.785), mentre *BrODE*, *FullVB* e *VeloVAE* mostrano valori negativi (-0.783, -0.831, -0.823). Questo conferma il backflow verso le popolazioni progenitrici osservato per la famiglia di modelli *VeloVAE*. Anche *cell2fate* presenta un valore negativo (-0.192), sebbene più contenuto; si nota anche come per questo metodo certe transizioni presentano valori positivi ed altre fortemente negativi, come Pre-endocrine $\rightarrow \beta$ (-0.548). L'alternanza tra segni opposti lungo transizioni diverse indica una direzione del flusso non uniforme piuttosto che un'inversione dell'intera traiettoria.

Metodo	Ductal → Ngn3 low EP	Ngn3 low EP → Ngn3 high EP	Ngn3 high EP → Pre-endocrine	Pre-endocrine → Alpha	Pre-endocrine → Beta	Pre-endocrine → Delta	Pre-endocrine → Epsilon
scVelo	0.193	0.881	0.648	0.651	0.755	0.274	-0.069
cell2fate	0.221	-0.192	0.059	-0.231	-0.548	0.269	-0.266
VeloVI	0.073	0.785	0.692	0.595	0.723	0.499	0.257
VeloVAE	-0.174	-0.823	0.560	0.351	0.504	-0.267	0.644
FullVB	-0.154	-0.831	0.138	0.515	0.648	0.064	0.502
BrODE	-0.210	-0.783	0.549	0.643	0.755	0.409	0.207

Tabella 4.3: Valori di CBDir ottenuti dai diversi metodi per ciascuna transizione biologica considerata nel dataset del pancreas.

La Tabella 4.4 riporta la coerenza intra-cluster per singola popolazione cellulare. *VeloVI* e *cell2fate* risultano i metodi con la coerenza locale più elevata tra le popolazioni, con valori quasi sempre superiori a 0.9. *VeloVAE* mostra invece la maggiore variabilità tra popolazioni, con valori compresi tra 0.531 (Alpha) e 0.861 (Pre-endocrine), a indicare una coerenza locale del campo di velocità meno uniforme rispetto agli altri metodi. *scVelo*, pur avendo il proprio valore più basso sulla popolazione Epsilon (0.600), mantiene per le restanti popolazioni valori

Metodo	Alpha	Beta	Delta	Ductal	Epsilon	Ngn3 high EP	Ngn3 low EP	Pre-endocrine
scVelo	0.802	0.847	0.921	0.757	0.600	0.827	0.782	0.848
cell2fate	0.979	0.987	0.981	0.906	0.975	0.896	0.896	0.982
VeloVI	0.974	0.971	0.968	0.946	0.961	0.944	0.944	0.967
VeloVAE	0.531	0.708	0.746	0.746	0.662	0.801	0.627	0.861
FullVB	0.766	0.950	0.973	0.903	0.942	0.837	0.860	0.873
BrODE	0.952	0.939	0.932	0.916	0.940	0.955	0.896	0.945

Tabella 4.4: Valori di ICCoh ottenuti dai diversi metodi per ciascun cluster del dataset del pancreas.

nell'intervallo tipico degli altri metodi (0.757–0.921). *cell2fate* ottiene invece valori elevati e omogenei su tutte le popolazioni, pur avendo mostrato un campo di velocità qualitativamente invertito: un'ulteriore conferma che l'ICCoh, misurando solo la coerenza locale tra vicini dello stesso cluster, non è di per sé sufficiente a rilevare inversioni sistematiche della direzione del flusso, e va quindi sempre interpretata congiuntamente alla CBDir. Valori elevati di ICCoh non implicano necessariamente la correttezza biologica perché la struttura del grafo k-NN e lo smoothing impiegato nel preprocessing possono produrre campi localmente molto regolari anche quando la direzione inferita non riflette correttamente la dinamica sottostante.

Un aspetto degno di nota riguarda la popolazione Ductal, nota in letteratura per il suo carattere ciclico piuttosto che direzionale verso un singolo destino: ci si aspetterebbe quindi una minore coerenza locale delle velocity in questa popolazione, ma per la maggior parte dei metodi le Ductal mostrano ICCoh elevato, paragonabile o superiore a quello delle popolazioni terminali, suggerendo che i modelli non stiano catturando esplicitamente la loro dinamica ciclica, producendo invece un segnale di velocity artificialmente lineare.

Metodo	Alpha	Beta	Delta	Epsilon
scVelo	0.950	0.943	0.921	0.927
cell2fate	0.855	0.844	0.895	0.889
VeloVI	0.884	0.893	0.841	0.847
VeloVAE	-0.524	-0.465	-0.681	-0.649
FullVB	0.638	0.673	0.428	0.474
BrODE	-0.524	-0.465	-0.681	-0.649

Tabella 4.5: Correlazione tra il tempo latente stimato e l'ordine biologico atteso lungo le diverse linee endocrine del dataset del pancreas.

La Tabella 4.5 conferma quanto osservato a livello aggregato dalle metriche precedenti e dai campi di velocità per quanto riguarda la correlazione tra tempo latente e ordine biologico atteso. *scVelo*, *VeloVI* ottengono correlazioni positive su tutte e quattro le linee endocrine. *VeloVAE* e *BrODE* mostrano invece correlazioni negative su ogni linea (da -0.681 a -0.465), identiche tra i due metodi poiché *BrODE* eredita il tempo latente di *VeloVAE* senza modificarlo. *FullVB* si colloca in posizione intermedia, con correlazioni positive ma più basse (0.428–0.673). *cell2fate* mostra un valore abbastanza alto nonostante abbia un'inversione nella parte finale della traiettoria. Le prime tappe del differenziamento (Ductal → Ngn3 low EP → Ngn3 high EP → Pre-endocrine) vengono però ordinate correttamente.

Dato il fenomeno di backflow marcato osservato è naturale chiedersi se tali risultati siano una caratteristica sistematica del modello su questo dataset o dipendano invece dalla particolare inizializzazione casuale utilizzata. Le sottosezioni seguenti approfondiscono rispettivamente questo aspetto per *cell2fate* e, in modo analogo, la sensibilità di *VeloVAE* rispetto agli iperparametri di addestramento.

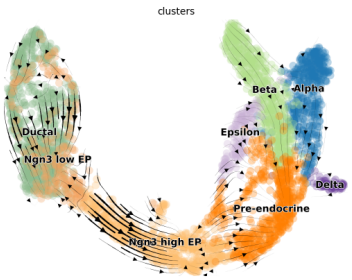
Sensibilità rispetto all'inizializzazione di *cell2fate*

Poiché l'inferenza di *cell2fate* è basata su un procedimento di ottimizzazione stocastica, è stata valutata la sensibilità del risultato rispetto all'inizializzazione casuale utilizzando tre diversi valori del seed (0, 123 e 2026), mantenendo invariati il preprocessing e tutti gli altri iperparametri.

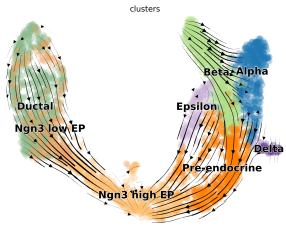
Seed	CBDir	ICCoh	CS graph	CS Pearson	Latent time corr.
0	-0.098	0.950	0.939	0.941	0.871
123	-0.193	0.720	0.726	0.733	0.688
2026	0.565	0.869	0.885	0.886	0.942

Tabella 4.6: Valori delle metriche di valutazione ottenuti da *cell2fate* sul dataset pancreatico al variare del seed di inizializzazione

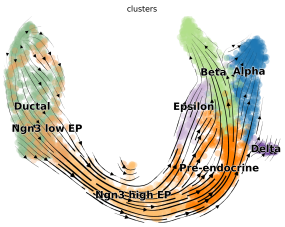
Come mostrato nella Figura 4.2, la scelta del seed modifica la densità delle linee di flusso e l'orientamento dei vettori. In particolare, nelle esecuzioni con seed 0 e soprattutto 123 si osservano velocità orientate dalle popolazioni endocrine mature verso la popolazione Pre-endocrine e gli stati progenitori. Con il seed 2026, invece, il campo di velocità ricostruisce correttamente la progressione dalle popolazioni Ngn3 low EP, passando per le Ngn3 high EP verso le cellule Pre-endocrine



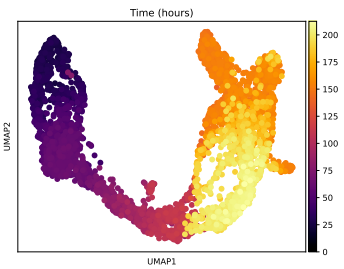
(a) Seed 0 – Velocity field



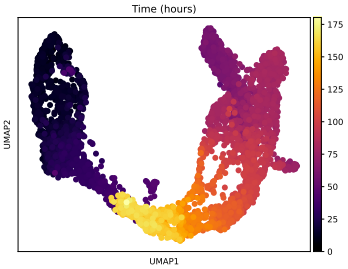
(b) Seed 123 – Velocity field



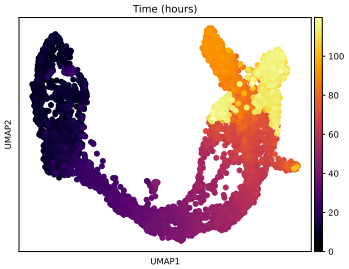
(c) Seed 2026 – Velocity field



(d) Seed 0 – Latent time



(e) Seed 123 – Latent time



(f) Seed 2026 – Latent time

Figura 4.2: Confronto delle RNA velocity e del tempo latente stimati da *cell2fate* sul dataset pancreatico utilizzando differenti seed.

e, successivamente, verso gli stati endocrini maturi. Il fenomeno di backflow osservato nelle altre esecuzioni non è quindi conservato al variare dell'inizializzazione, evidenziando una rilevante sensibilità del modello alla scelta del seed.

L'analisi del tempo latente evidenzia tuttavia una parziale incongruenza nell'ordinamento temporale degli stati maturi. In particolare, alle cellule α ed ϵ vengono assegnati tempi leggermente successivi rispetto alle cellule β, nonostante queste ultime siano attese nella fase più tardiva del processo di differenziamento. Pertanto, sebbene l'esecuzione con seed 2026 produca un campo di velocità complessivamente coerente con la traiettoria nota, l'ordinamento temporale delle popolazioni mature non risulta pienamente concorde con le conoscenze biologiche.

La Tabella 4.6 conferma quantitativamente quanto osservato nella Figura 4.2: la scelta del seed non produce semplici oscillazioni numeriche attorno a un valore stabile, ma risultati qualitativamente diversi. Il seed 0 ottiene la ICCoh e i Consistency Score più elevati fra le tre esecuzioni, ma una CBDir negativa (-0.098), a conferma che un'elevata coerenza locale del campo può coesistere con una direzione complessivamente errata rispetto alla traiettoria attesa. Il seed 123 presenta il quadro peggiore su quasi tutte le metriche, con una CBDir ancora negativa (-0.193) e la correlazione temporale più bassa (0.688). Solo il seed 2026 ottiene una

CBDir positiva (0.565) e la correlazione temporale più elevata (0.942), risultati coerenti con il campo di velocità corretto osservato in Figura 4.2c.

Il backflow è molto marcato e coinvolge più popolazioni simultaneamente; questo può essere attribuito alla struttura modulare del modello. Il tasso di trascrizione di ciascun gene non è stimato in modo indipendente, ma è il risultato di una somma di M termini ($\alpha_{mg}(t)$) condivisi tra geni e governati da tempi di accensione t_m^{on} e spegnimento t_m^{off} comuni. Questa condivisione è resa ancora più delicata dal fatto che il tempo cellulare globale T_c non è osservato né fissato, ma generato a sua volta da una gerarchia di prior: $T_c = t_c \cdot T_{\max}$, con $t_c \sim \text{Gamma}(t_{\text{loc}}, t_{\text{scale}})$ e sia t_{loc} sia t_{scale} variabili casuali a loro volta, mentre gli offset temporali di attivazione di ciascun modulo, $\delta\tau_m^{\text{ON}}$, sono anch'essi distribuiti come Gamma in funzione della scala complessiva $T_{\max}$ e del numero di moduli M.

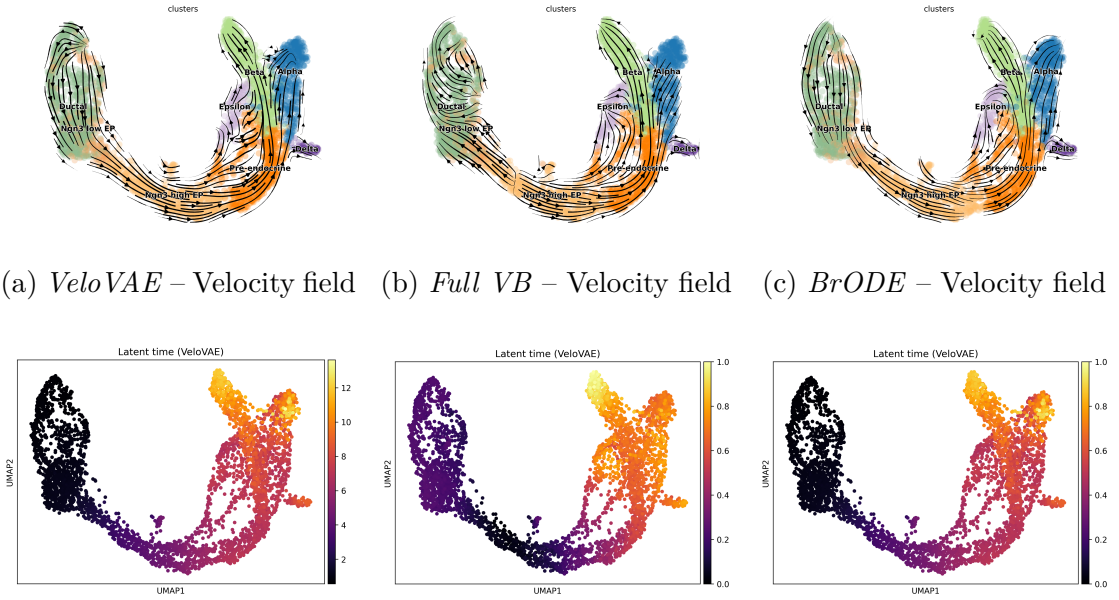
Questi risultati evidenziano che l'inferenza di *cell2fate* sul dataset pancreatico non risulta sufficientemente stabile rispetto all'inizializzazione casuale. A parità di dati e iperparametri, seed differenti producono sia traiettorie corrette sia fenomeni di backflow.

Analisi di sensibilità agli iperparametri di *VeloVAE*

Per *VeloVAE* è stata condotta un'analisi di sensibilità rispetto agli iperparametri di addestramento. Si è provato a modificare la configurazione di default che produceva un'inversione del campo di velocità. Il modello è stato quindi nuovamente addestrato modificando i learning rate delle diverse componenti dell'ottimizzazione. Impostando il learning rate generale a $\eta = 10^{-4}$, il learning rate relativo alla componente ODE a $\eta_{ODE} = 50\eta = 5 \times 10^{-3}$ e il learning rate della fase di post-training a $\eta = 10^{-4}$, il modello standard restituisce un campo di velocità coerente con la linea attesa, in Figura 4.3. Lo stesso accade per *BrODE*.

Lo stesso miglioramento non si osserva tuttavia per il modello *FullVB*, che con questa configurazione mantiene un andamento molto simile alla Figura 4.1h. Questo suggerisce che la correzione ottenuta variando i soli learning rate generali non sia sufficiente per *FullVB*. La scelta di intervenire sui learning rate relativi delle diverse componenti dell'ottimizzazione, aumentando η_{ODE} e diminuendo η, non è stata una scelta arbitraria ma segue un'indicazione esplicita fornita dagli stessi sviluppatori di *VeloVAE*, secondo cui questo intervento è il rimedio generale raccomandato quando il modello mostra un comportamento scorretto con la configurazione di default.

È stato quindi osservato che il risultato dipende sensibilmente dagli iperparametri utilizzati durante l'addestramento. La scelta di aumentare sensibilmente il rapporto tra il learning rate dei parametri cinetici e quello della rete neurale ($\eta_{\text{ODE}} = 50\,\eta$, contro un rapporto di default di circa $2.5\,\eta$ nell'implementazione



(a) *VeloVAE* – Velocity field (b) *Full VB* – Velocity field (c) *BrODE* – Velocity field

(d) *VeloVAE* – Latent time (e) *Full VB* – Latent time (f) *BrODE* – Latent time

Figura 4.3: Confronto dei campi di velocità e dei tempi latenti stimati dalle tre varianti di *VeloVAE* con modifiche agli iperparametri sul dataset pancreatico.

originale) trova una motivazione nella descrizione che gli stessi autori di *VeloVAE* forniscono in merito al *vanishing velocity problem*: in assenza di un adeguato bilanciamento tra la velocità di adattamento dei tassi cinetici e quella della rappresentazione latente appresa dalla rete neurale, l'ottimizzazione può convergere verso una configurazione dei tassi che produce una dinamica degenere o, come osservato nella configurazione di default in questo lavoro, invertita rispetto alla direzione attesa. Aumentando il learning rate relativo dei parametri cinetici, questi ultimi vengono forzati ad adattarsi più rapidamente allo stato latente stimato dalla rete.

Gli stessi autori riconoscono peraltro che l'individuazione di iperparametri di default universalmente validi resta un problema aperto nell'implementazione del metodo.

La forte dipendenza dai parametri di addestramento rappresenta un limite pratico del metodo. Piccole variazioni della configurazione possono infatti modificare significativamente le traiettorie inferite, fino a produrre inversioni del campo di velocità. Di conseguenza, in applicazioni esplorative, nelle quali la dinamica biologica non è nota a priori, risulta difficile stabilire se la traiettoria ricostruita sia plausibile.

4.3.2 Neurogenesi del giro dentato

In questa sezione vengono presentati risultati analoghi anche per il dataset della neurogenesi del giro dentato.

Metodo	CBDir	ICCoh	CS graph	CS Pearson	Latent time corr.
scVelo	0.295	0.787	0.762	0.757	0.656
cell2fate	0.429	0.708	0.638	0.635	0.421
VeloVI	0.554	0.861	0.827	0.823	-0.186
VeloVAE	0.249	0.879	0.779	0.779	0.177
FullVB	0.237	0.906	0.864	0.865	0.212
BrODE	0.293	0.940	0.909	0.909	0.177

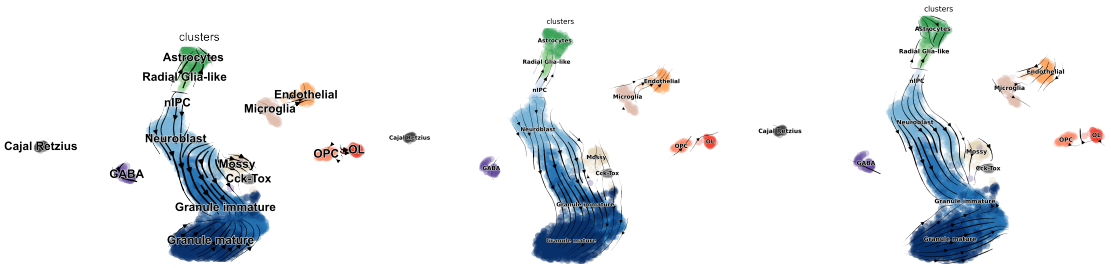
Tabella 4.7: Risultati delle metriche di valutazione per i metodi di RNA velocity applicati al dataset del giro dentato.

Dalla Tabella 4.7 emerge una maggiore concordanza tra i metodi rispetto a quanto osservato per il dataset del pancreas, con CBDir generalmente più basse. La Figura 4.4 mostra come tutti gli approcci ricostruiscano una traiettoria principale che parte da nIPC, attraversa la popolazione Neuroblast fino a raggiungere le cellule Granule immature, in accordo con il processo di maturazione delle cellule granulari del giro dentato.

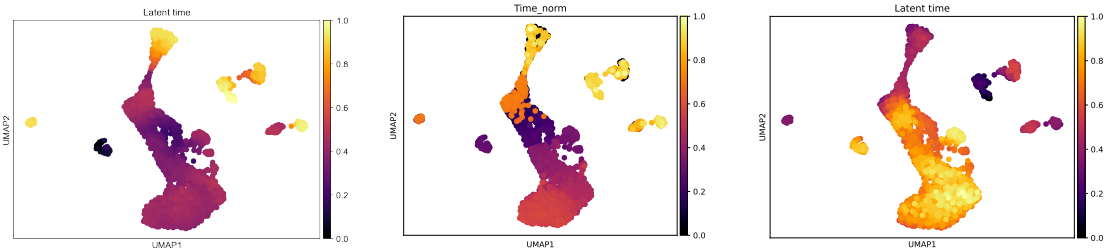
scVelo, *VeloVI* e *VeloVAE* mostrano un'inversione del campo di velocità in corrispondenza delle cellule Granule mature, con le frecce orientate prevalentemente dalle cellule granulari mature verso quelle immature: questo si riflette in una CBDir aggregata modesta per tutti.

VeloVI ottiene la CBDir più alta (0.554) ma una correlazione temporale negativa, mentre *scVelo*, *VeloVAE* (tutte le varianti) si collocano su valori più bassi (tra 0.237 e 0.295). *BrODE* mantiene tuttavia il valore di ICCoh più elevato (0.940), tuttavia una coerenza locale del campo alta non implica necessariamente una direzione complessivamente più corretta.

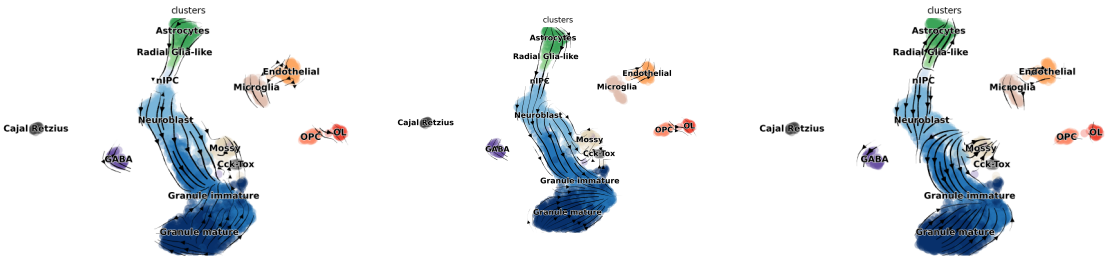
cell2fate ottiene una CBDir tra le più alte (0.429), infatti sembra ricostruire bene le linee di velocità. La scomposizione per singola transizione, discussa più avanti, (Tabella 4.8), evidenzia come questo sia il risultato del corretto andamento su quasi tutte le transizioni considerate (0.634, 0.559, 0.431, 0.588), anche verso le Granule mature, con la sola eccezione della transizione nIPC $\rightarrow$ Neuroblast, dove il valore è debolmente negativo. Presenta tuttavia valori di Consistency Score e ICCoh più bassi (0.635, 0.638, 0.708) rispetto agli altri metodi, mentre la correlazione temporale (0.421) è la seconda più elevata.



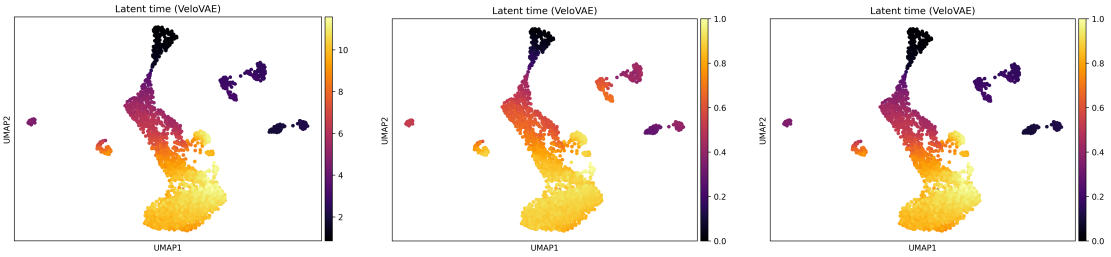
(a) *scVelo* – Velocity field (b) *cell2fate* – Velocity field (c) *VeloVI* – Velocity field



(d) *scVelo* – Latent time (e) *cell2fate* – Latent time (f) *VeloVI* – Latent time



(g) *VeloVAE* – Velocity field (h) *Full VB* – Velocity field (i) *BrODE* – Velocity field



(j) *VeloVAE* – Latent time (k) *Full VB* – Latent time (l) *BrODE* – Latent time

Figura 4.4: Confronto delle stime di RNA velocity e del tempo latente ottenute con *scVelo*, *cell2fate*, *VeloVI* e *VeloVAE* (standard, Full VB, BrODE) sul dataset del giro dentato.

Le due versioni di Consistency Score restano molto vicine tra loro per ogni metodo, confermando come per il pancreas che la scelta della statistica (cosine similarity/Pearson) pesi poco sul risultato finale. Le tre formulazioni di *VeloVAE*

producono risultati numericamente simili tra loro, un quadro confermato dalla Figura 4.4: l'unica differenza qualitativa rilevante riguarda la direzione del flusso nella regione degli Astrocytes, dove sia la formulazione standard sia *Full VB* mostrano un flusso invertito, mentre *BrODE* ricostruisce correttamente la direzione attesa. Coerentemente, *BrODE* ottiene una CBDir leggermente più alta tra le tre varianti (0.293 contro 0.237 e 0.249). *BrODE* riceve da *VeloVAE* standard solo il tempo latente e lo stato cellulare, ma ristima da zero i parametri cinetici ($\alpha_y, \beta_y, \gamma_y$ per ciascun tipo cellulare y) tramite regressione, invece di riutilizzare direttamente la velocity del modello di partenza: poiché ogni tipo cellulare condivide pochi parametri, anziché dipendere da una rete neurale complessa come nel modello standard, l'errore locale osservato negli Astrocytes non si propaga alla stima di *BrODE*. Il fatto che la correlazione temporale sia identica tra *VeloVAE* standard e *BrODE* (0.177 in entrambi i casi) è coerente con il fatto che *BrODE* eredita direttamente il tempo latente stimato da *VeloVAE* standard senza modificarlo: la correzione osservata nel campo di velocità riguarda quindi esclusivamente la stima dei tassi cinetici, non l'ordinamento temporale delle cellule. *FullVB* mostra invece una correlazione leggermente più alta (0.212), la migliore tra le tre varianti, sebbene tutte restino nettamente inferiori a quella di *scVelo* (0.656), che risulta il metodo con il tempo latente più accurato sull'intero dataset.

Metodo	**nIPC → Neuroblast**	**Neuroblast → Granule immature**	**Granule immature → Granule mature**	**OPC → OL**	**Radial Glia-like → Astrocytes**
scVelo	0.911	-0.096	-0.074	-0.109	0.841
cell2fate	-0.065	0.634	0.559	0.431	0.588
VeloVI	0.899	0.595	-0.154	0.603	0.829
VeloVAE	0.919	0.619	-0.010	0.553	-0.838
FullVB	0.930	0.599	-0.054	0.554	-0.846
BrODE	0.915	0.666	-0.062	0.773	-0.826

Tabella 4.8: Valori di CBDir ottenuti dai diversi metodi per ciascuna transizione biologica considerata nel dataset del giro dentato.

Scomponendo il valore aggregato di CBDir nelle singole transizioni biologiche (Tabella 4.8), emergono le seguenti considerazioni. La transizione Granule immature → Granule mature conferma il backflow discusso qualitativamente per la maggior parte dei metodi, con valori negativi o vicini allo zero, con la sola eccezione di *cell2fate*, che su questa transizione ottiene il valore più alto (0.559), come era emerso nella Figura 4.4b.

Un altro aspetto rilevante riguarda la transizione Radial Glia-like → Astrocytes: *scVelo* (0.841) e *VeloVI* (0.829) ottengono valori fortemente positivi, e anche *cell2fate* (0.588) risulta coerente con la direzione attesa, mentre l'intera famiglia *VeloVAE* mostra valori negativi (-0.838, -0.846, -0.826), indicando un'inversione sistematica su questa linea, indipendente dalla variante considerata. Questo

caso evidenzia come le metriche quantitative permettano di rilevare fenomeni di backflow localizzati specialmente in presenza di cluster compressi nello spazio di riduzione dimensionale.

Si osservi inoltre che in alcune delle cinque transizioni presenti c'è almeno un metodo in disaccordo con tutti gli altri: *cell2fate* nella transizione nIPC → Neuroblast e anche Granule immature → Granule mature; *scVelo* in OPC → OL e in Neuroblast → Granule immature.

Va infine notato che *VeloVI*, pur ottenendo la CBDir aggregata più elevata, mostra la correlazione tra tempo latente e ordine biologico peggiore tra tutti i metodi (-0.186, l'unico valore negativo della Tabella 4.7): come già osservato per il pancreas, una direzione locale del flusso corretta non garantisce un ordinamento temporale complessivamente coerente con la traiettoria nota.

La Tabella 4.9 riporta la coerenza intra-cluster per singola popolazione cellulare. Solo *cell2fate* mostra sistematicamente i valori più bassi nelle popolazioni progenitrici (Neuroblast, nIPC), coerentemente con l'instabilità osservata per questo metodo proprio su queste transizioni nella Tabella 4.8. In *FullVB* il pattern è invece invertito: i valori minimi si osservano in Cck-Tox, Mossy e GABA, mentre le popolazioni progenitrici restano su valori elevati. Negli altri metodi (*BrODE*, *scVelo*, *VeloVAE*, *VeloVI*) il valore minimo della riga appartiene quasi sempre a una popolazione matura o secondaria (Microglia, Granule mature, Astrocytes, Endothelial), non a un progenitore. La minore coerenza delle popolazioni progenitrici non è quindi una caratteristica sistematica del giro dentato, ma dipende fortemente dal metodo considerato. La famiglia *VeloVAE* sulle popolazioni Radial Glia-like e Astrocytes ottiene alcuni dei valori di ICCoh più alti dell'intera riga nonostante sia proprio su questa transizione che la Tabella 4.8 rileva un'inversione. Come osservato per cell2fate sul dataset pancreatico, anche qui un'elevata ICCoh non garantisce la correttezza direzionale. Il fenomeno, limite della metrica, non è quindi un caso isolato legato a un singolo dataset.

Metodo	Astrocytes	Cajal Retzius	Cck-Tox	Endothelial	GABA	Granule immature	Granule mature	Microglia	Mossy	Neuroblast	Radial Glia-like	nIPC
scVelo	0.768	0.992	0.898	0.887	0.851	0.773	0.764	0.699	0.813	0.702	0.741	0.711
cell2fate	0.780	0.496	0.680	0.720	0.810	0.770	0.851	0.916	0.714	0.479	0.715	0.496
VeloVI	0.805	0.945	0.922	0.815	0.878	0.830	0.812	0.845	0.840	0.850	0.865	0.875
VeloVAE	0.901	0.993	0.990	0.743	0.800	0.797	0.704	0.933	0.822	0.808	0.967	0.863
FullVB	0.997	0.958	0.755	0.923	0.762	0.867	0.807	0.968	0.758	0.931	0.990	0.966
BrODE	0.988	0.934	0.957	0.981	0.877	0.973	0.966	0.996	0.930	0.960	0.952	0.797

Tabella 4.9: Valori di ICCoh ottenuti dai diversi metodi per ciascun cluster del dataset del giro dentato.

La Tabella 4.10 scompone la correlazione tra tempo latente e ordine biologico atteso per le tre principali linee di differenziamento del dataset. Sulla linea granulare (Granule lineage) tutti i metodi ottengono correlazioni positive ma piuttosto

Metodo	Granule lineage	Oligodendrocyte lineage	Astrocyte lineage
scVelo	0.318	0.866	0.784
cell2fate	0.579	0.812	-0.127
VeloVI	0.202	-0.866	0.106
VeloVAE	0.462	0.846	-0.777
FullVB	0.554	0.866	-0.784
BrODE	0.462	0.846	-0.777

Tabella 4.10: Correlazione tra il tempo latente stimato e l'ordine biologico atteso lungo le diverse linee di differenziamento del dataset del giro dentato.

contenute, *cell2fate* ottiene 0.579, leggermente più alto perché riesce a catturare anche la transizione finale delle Granule mature.

Per quanto riguarda quella oligodendrocitaria la maggior parte dei metodi mostra correlazioni elevate (0.812–0.866), con l'eccezione di *VeloVI*, che ottiene un valore negativo (-0.866). Questo valore fortemente negativo spiega, almeno in parte, il valore di -0.186 osservato per VeloVI in Tabella 4.7: la correlazione aggregata corrisponde infatti alla media delle tre linee.

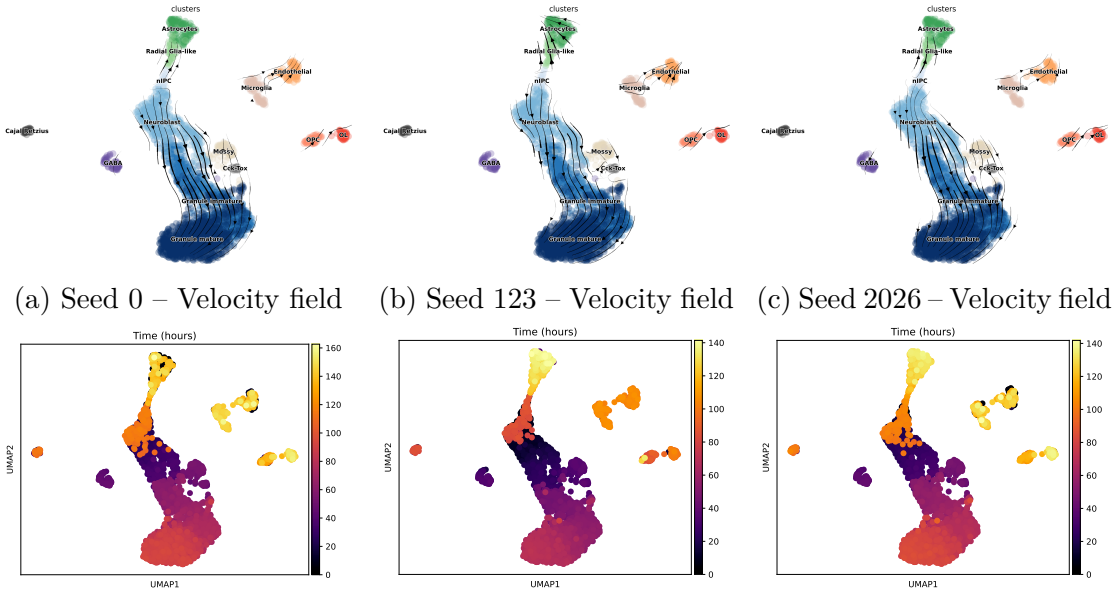
Sull'ultima linea, quella astrocitaria, si osserva la stessa separazione già evidenziata dalla CBDir: l'intera famiglia *VeloVAE* mostra correlazioni negative e pressoché identiche.

Sensibilità di *cell2fate* all'inizializzazione

Anche per il dataset del giro dentato è stata valutata la sensibilità dell'inferenza di *cell2fate* rispetto all'inizializzazione casuale, utilizzando gli stessi tre seed già adottati per il pancreas (0, 123 e 2026). La Tabella 4.11 riassume le metriche ottenute per ciascun seed, mentre la Figura 4.5 mostra i corrispondenti campi di velocità e tempi latenti.

Seed	CBDir	ICCoh	CS graph	CS Pearson	Latent time corr.
0	0.429	0.708	0.638	0.635	0.421
123	0.529	0.726	0.642	0.631	0.671
2026	0.387	0.733	0.683	0.678	0.684

Tabella 4.11: Metriche di valutazione ottenute da *cell2fate* sul dataset del giro dentato al variare del seed di inizializzazione.



(a) Seed 0 – Velocity field (b) Seed 123 – Velocity field (c) Seed 2026 – Velocity field

(d) Seed 0 – Latent time (e) Seed 123 – Latent time (f) Seed 2026 – Latent time

Figura 4.5: Confronto delle RNA velocity e del tempo latente stimati da *cell2fate* sul dataset del giro dentato utilizzando differenti seed.

A differenza di quanto osservato per il pancreas, in questo caso le tre esecuzioni risultano sostanzialmente coerenti tra loro senza differenze rilevanti nella direzione del campo di velocità al variare del seed.

I tre seed producono valori di CBDir aggregata positiva (0.429, 0.529, 0.387), senza l'inversione di segno marcata riscontrata sul dataset pancreatico. Anche l'ICCoh (0.708–0.733) e i due Consistency Score restano relativamente stabili tra le tre esecuzioni. Dalla Figura 4.5 emergono alcune osservazioni. La transizione Radial Glia-like → Astrocytes, che nella famiglia *VeloVAE* risultava sistematicamente invertita, è invece positiva e concorde su tutti e tre i seed, indicando che *cell2fate* riesce a ricostruirla indipendentemente dall'inizializzazione. L'unica transizione che mostra un comportamento incerto è nIPC → Neuroblast, suggerendo una minore stabilità nella fase iniziale della linea granulare. L'analisi del tempo latente conferma questa maggiore stabilità rispetto al pancreas: i tre seed producono un ordinamento temporale delle popolazioni positivo, senza le discrepanze osservate per le popolazioni endocrine mature del pancreas.

È opportuno precisare che questa osservazione si basa su un numero limitato di inizializzazioni e non consente quindi di concludere in modo generale che l'inferenza di *cell2fate* sul dataset del giro dentato sia robusta rispetto al seed; è possibile che inizializzazioni diverse da quelle testate producano risultati differenti, come osservato per il pancreas.

L'analisi di sensibilità rispetto agli iperparametri, condotta analogamente a quanto descritto in precedenza per *VeloVAE*, non viene riportata, poiché non ha evidenziato differenze sostanziali rispetto ai risultati già osservati.

4.3.3 Confronto tra *scVelo* e *VeloVI*

Nel confronto precedente, *scVelo* e *VeloVI* hanno prodotto stime della dinamica complessivamente molto simili. Essi condividono lo stesso modello matematico alla base e stimano puntualmente i tassi cinetici gene-specifici della dinamica. Tuttavia *VeloVI* introduce una variabile latente a bassa dimensionalità che collega i geni tra loro e riformula l'inferenza mediante VAEs. Questa formulazione permette di ottenere una distribuzione a posteriori delle velocità. Gli autori introducono inoltre due quantità per caratterizzare l'incertezza associata alle stime di RNA velocity: l'**intrinsic uncertainty**, che descrive la variabilità nella direzione della velocità stimata in ciascuna cellula, e l'**extrinsic uncertainty**, che tiene conto anche della variabilità degli stati cellulari futuri predetti attraverso la matrice di transizione. Riportiamo queste analisi per valutare l'incertezza del campo di velocità poiché il modello EM di *scVelo* non fornisce una misura della variabilità associata alle stime. Si affianca a questa analisi anche i *Phase Portraits* che rappresentano la relazione tra le abbondanze di RNA spliced e unspliced. Un'analisi dei singoli geni consente di esaminare direttamente le dinamiche trascrizionali che contribuiscono alla costruzione del campo di velocità. La forma del *Phase Portraits*, in particolare la curvatura rispetto alla linea di stato stazionario, permette di distinguere le fasi di induzione e repressione di un gene. Bergen et al. [2021] sottolineano che una curvatura poco evidente e la presenza di regimi multipli riducono la capacità di identificare correttamente la direzione della dinamica.

Intrinsic uncertainty

Per ogni cellula i siano $v_i^{(1)}$, $v_i^{(2)}$, $\dots$, $v_i^{(L)}$ L campioni a posteriori della sua velocità Sia inoltre $\bar{v}_i$ la loro media. L'intrinsic uncertanty è

$$\sigma^2_{i,\text{ext}} = \frac{1}{L-1}\sum_{l=1}^{L}\left(\frac{v_i^{(l)}\cdot\bar{v}_i}{\left\|v_i^{(l)}\right\|\|\bar{v}_i\|} - \frac{1}{L}\sum_{j=1}^{L}\frac{v_i^{(j)}\cdot\bar{v}_i}{\left\|v_i^{(j)}\right\|\|\bar{v}_i\|}\right)^2$$

che corrisponde alla varianza della cosine similarity tra la media a posteriori e l-esimo campione di velocity.

Descrive quanto è incerta la direzione del vettore di RNA velocity di una cellula misurando la variabilità dei vettori di velocity campionati dalla distribuzione a

posteriori rispetto alla loro direzione media. Maggiore è il suo valore maggiore sarà la variabilità e quindi minore l'affidabilità della direzione stimata per la cellula.

Extrinsic uncertainty

Per ogni campione l della distribuzione a posteriori della velocity, sia

$$V^{(l)} = \left(v_1^{(l)}, \ldots, v_C^{(l)}\right)$$

l'insieme dei vettori di velocity delle C cellule e sia $\tilde{\pi}^{(l)}$ la matrice di transizione, costruita come nel post-processing di *scVelo*. Lo stato cellulare futuro atteso viene calcolato come

$$\tilde{F}^{(l)} = \tilde{\pi}^{(l)} S,$$

dove S è la matrice delle abbondanze di RNA spliced. La i-esima riga di $\tilde{F}^{(l)}$, indicata con $\tilde{f}_i^{(l)}$, rappresenta lo stato futuro atteso della cellula i, ottenuto a partire dal l-esimo campione. L'extrinsic uncertainty si ottiene come l'intrinsic uncertainty, cioè calcolando la varianza della similarità coseno tra ciascuno stato futuro campionato e lo stato futuro medio:

$$\sigma^2_{i,\text{ext}} = \frac{1}{L-1} \sum_{l=1}^{L} \left(\frac{\tilde{f}_i^{(l)} \cdot \bar{f}_i}{\left\|\tilde{f}_i^{(l)}\right\| \left\|\bar{f}_i\right\|} - \frac{1}{L} \sum_{j=1}^{L} \frac{\tilde{f}_i^{(j)} \cdot \bar{f}_i}{\left\|\tilde{f}_i^{(j)}\right\| \left\|\bar{f}_i\right\|} \right)^2 .$$

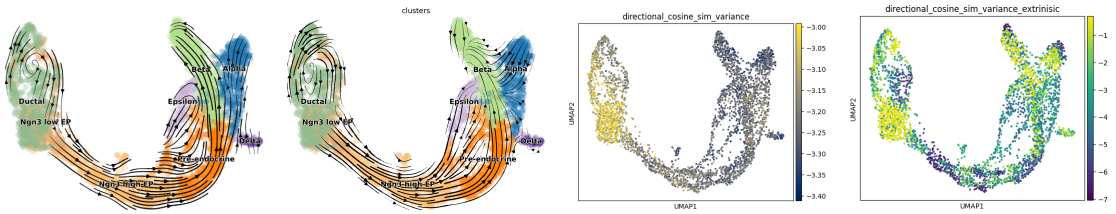
dove $\bar{f}_i$ è lo stato futuro medio.

Non misura quindi quanto cambiano i vettori di velocità, ma quanto questa variabilità si ripercuote sugli stati cellulari futuri. Valori elevati indicano che diversi campioni della distribuzione a posteriori della velocity conducono a stati cellulari futuri differenti, e quindi a una maggiore incertezza nelle transizioni predette.

Endocrinogenesi pancreatica

In Figura 4.6c l'intrinsic uncertainty sembra maggiore soprattutto nella regione corrispondente alle cellule Ductal e Ngn3 low EP: questo suggerisce che i campioni a posteriori presentino direzioni meno concordi. Nelle popolazioni intermedie e mature la direzione della velocity risulta più stabile.

Nell'UMAP dell'extrinsic uncertainty, Figura 4.6d, si evince una distribuzione più eterogenea e una maggiore variazione. I valori più elevati si concentrano nella popolazione Ductal, coerentemente con l'elevata incertezza intrinseca, e nelle popolazioni α e β. Questo è attribuito al fatto che sono popolazioni terminali, prive di stati successori. Questo pattern risulta coerente con quanto riportato dagli autori di VeloVI Gayoso et al. [2024]. Al contrario, popolazioni appartenenti a



(a) *scVelo* (b) *VeloVI* (c) Intrinsic uncertainty (d) Extrinsic uncertainty

Figura 4.6: Dinamiche inferite da *scVelo* e *VeloVI* nel dataset del pancreas e distribuzione dell'intrinsic e dell'extrinsic uncertainty stimate da *VeloVI*.

stati di transizione hanno bassa incertezza in entrambe le metriche, poichè hanno successori ben definiti.

Le Figure 4.7 e 4.8 mostrano i *Phase portraits* di alcuni geni: *Pcsk2*, *Nnat* e *Cpe*. Per *Cpe* le traiettorie ricostruite dai due metodi sono molto simili. Per *Pcsk2* e, soprattutto, per *Nnat*, *VeloVI* stima invece una traiettoria più ampia, con un ramo di repressione che attraversa una regione del *phase portrait* scarsamente popolata. *scVelo* invece restituisce una traiettoria concentrata in prossimità delle osservazioni. I *Phase portraits* di *VeloVI* mostrano somiglianze nella forma delle traiettorie tra i diversi geni. Questo comportamento può essere ricondotto alla struttura latente del modello, che collega i geni attraverso una rappresentazione condivisa della cellula, favorendo una maggiore regolarità. Il confronto mostra quindi come le diverse procedure di inferenza possano condurre a ricostruzioni gene-specifiche differenti.

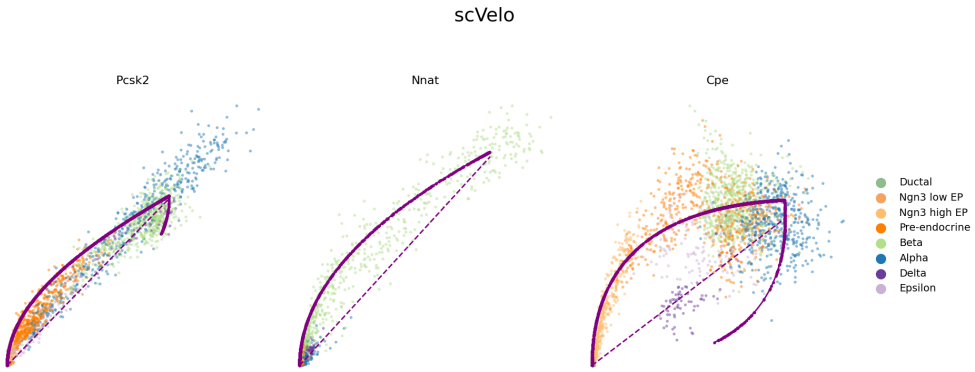


Figura 4.7: *Phase portraits* dei geni *Pcsk2*, *Nnat* e *Cpe* stimati da *scVelo*.

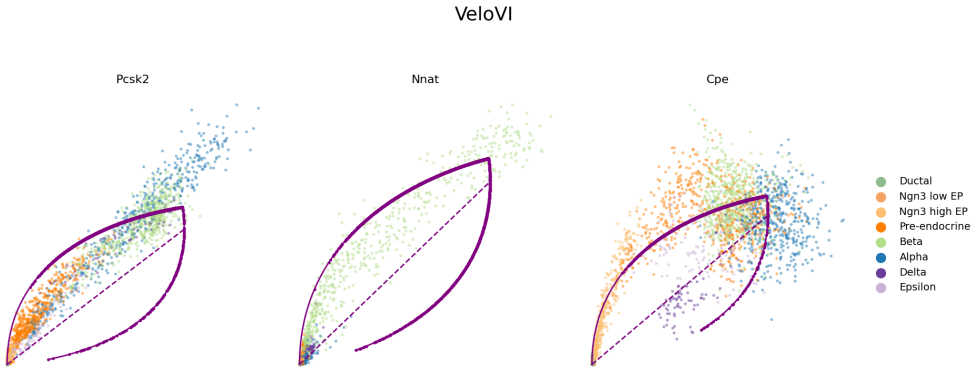
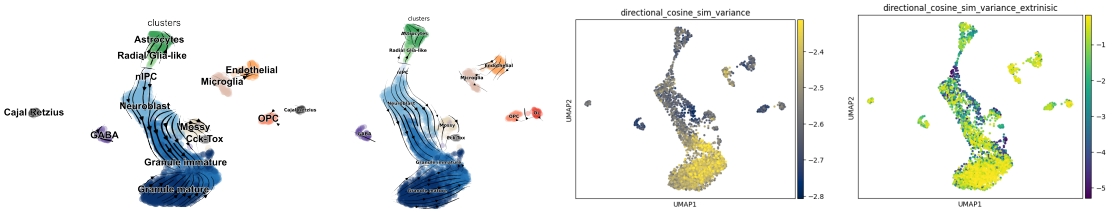


Figura 4.8: *Phase portraits* dei geni *Pcsk2*, *Nnat* e *Cpe* stimati da *VeloVI*.

Neurogenesi del giro dentato

L'intrinsic uncertainty, Figura 4.9c, presenta valori più elevati soprattutto nella regione inferiore, corrispondente alle Granule mature e Granule immature, dove il campo di velocità mostra un backflow. Risulta più contenuta nelle regioni corrispondenti ai progenitori. Anche per questo dataset l'extrinsic uncertainty, Figura 4.9d, presenta una distribuzione più eterogenea: sono presenti valori elevati anche per le popolazioni isolate, oltre che per le popolazioni terminali. Lungo la traiettoria principale le popolazioni progenitrici e transitorie mostrano complessivamente una maggiore stabilità delle transizioni rispetto alle popolazioni granulari. Il fatto che entrambe le uncertainty mostrino valori elevati proprio nelle regioni in cui il campo di velocità risulta biologicamente poco plausibile suggerisce che queste stime siano meno affidabili.



(a) *scVelo* (b) *VeloVI* (c) Intrinsic uncertainty (d) Extrinsic uncertainty

Figura 4.9: Dinamiche inferite da *scVelo* e *VeloVI* nel dataset del giro dentato e distribuzione dell'intrinsic e dell'extrinsic uncertainty stimate da *VeloVI*.

Per il dataset del giro dentato sono stati selezionati geni discussi in letteratura per le dinamiche di neurogenesi: *Fam155a*, *Hn1* e *Ppp3ca*.

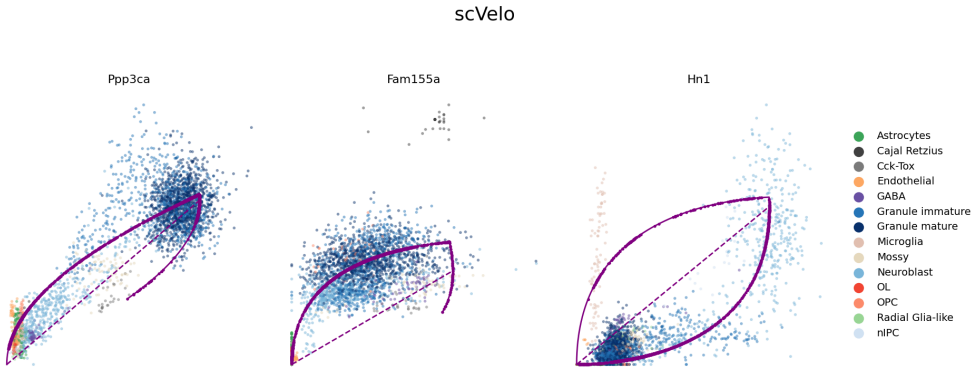


Figura 4.10: *Phase portraits* dei geni *Ppp3ca*, *Fam155a* e *Hn1* stimati da *scVelo* nel dataset dentate gyrus.

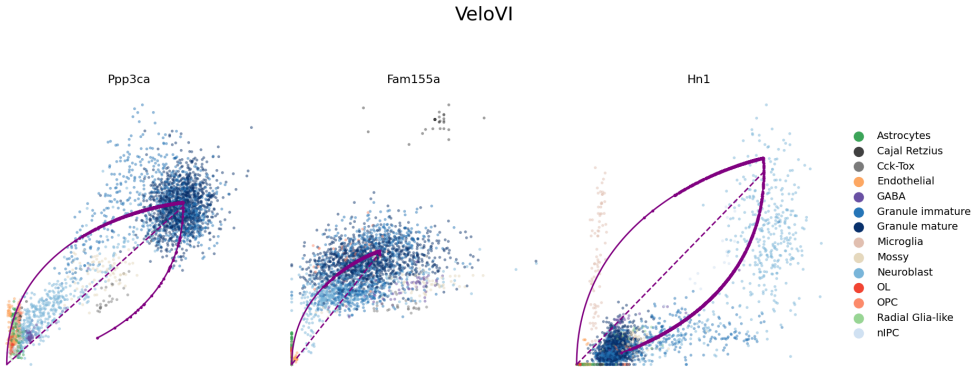


Figura 4.11: *Phase portraits* dei geni *Ppp3ca*, *Fam155a* e *Hn1* stimati da *VeloVI* nel dataset dentate gyrus.

Nel dataset dentate gyrus emergono differenze più marcate tra i due modelli, in Figura 4.10 e 4.11. Per *Ppp3ca*, le osservazioni di RNA formano una nube densa e continua, *VeloVI* produce tuttavia una curva leggermente più ampia, mentre quella di *scVelo* rimane concentrata nella regione più densamente popolata. Per *Fam155a*, la distribuzione delle cellule è più dispersa (si noti la presenza delle cellule stazionarie Cajal Retzius, che non appartengono alla traiettoria principale). *VeloVI* stima una traiettoria localizzata nella regione centrale dei dati, mentre *scVelo* descrive una curva più estesa. Infine, per *Hn1* entrambi i metodi ricostruiscono una traiettoria simile. *Hn1* è un esempio di gene con regimi cinetici multipli (si noti ad esempio la retta relativa alle cellule Microglia). La concordanza tra *scVelo* e *VeloVI* in questo caso non è quindi necessariamente indice di una stima corretta.

4.4 Conclusioni

L'obiettivo di questo lavoro è stato confrontare diversi metodi di RNA velocity (*scVelo*, *cell2fate*, *VeloVI* e *VeloVAE*, nelle varianti standard, *FullVB* e *BrODE*). In particolare, ci si è chiesti se le criticità osservate nell'inferenza siano riconducibili principalmente alle procedure algoritmiche oppure alle assunzioni del modello matematico alla base. È stata valutata sia la capacità di ricostruire le traiettorie cellulari mediante la visualizzazione del campo di velocità sia l'accuratezza delle stime attraverso metriche che valutavano la correttezza delle transizioni, la coerenza locale della velocità e l'ordinamento temporale delle cellule. Le metriche quantitative introdotte (CBDir, ICCoh, i due Consistency Score, la correlazione tra tempo latente e ordine biologico) non hanno individuato un metodo uniformemente superiore agli altri. Le diverse metriche valutano infatti proprietà differenti del campo di velocità e possono fornire indicazioni non concordi: un'elevata coerenza tra velocity di cellule vicine non garantisce che il flusso sia orientato correttamente. Più volte, per entrambi i dataset analizzati, un'elevata coerenza locale è stata osservata anche in presenza backflow, con un campo di velocità orientato in direzione opposta rispetto alla traiettoria biologica attesa. L'inversione del flusso è stata osservata, in misura diversa, in tutti i metodi testati: su più transizioni simultaneamente in *VeloVAE* e *cell2fate* sul pancreas; più localmente in *scVelo* e *VeloVI* sulle Granule immature del giro dentato. Il fatto che fenomeni di *backflow* emergano, seppur con caratteristiche differenti, sia nei modelli basati su inferenza EM sia negli approcci bayesiani e basati su reti neurali suggerisce che il problema non sia legato alla procedura di inferenza, ma sia, almeno in parte, riconducibile alla formulazione sottostante o alla presenza di un segnale cinetico debole che non consente di distinguere chiaramente le fasi di induzione e repressione previste dal modello. Le analisi di sensibilità hanno evidenziato inoltre un problema di stabilità. Per *cell2fate*, modificando il seed sono state ottenute traiettorie qualitativamente differenti sullo stesso dataset, ottenendo sia backflow sia ricostruzioni attese. *VeloVAE* ha mostrato una forte dipendenza dagli iperparametri di ottimizzazione: modificando i learning rate della rete neurale è stato possibile correggere alcune delle inversioni osservate con la configurazione di default. In contesti in cui la direzione attesa non è nota, questi risultati rappresentano un limite per l'affidabilità dell'RNA velocity come strumento esplorativo. La capacità di quantificare l'incertezza – assente in *scVelo* e presente in *cell2fate*, *FullVB* e *VeloVI* – consente di individuare le stime meno affidabili. Nel dataset pancreatico, *scVelo* e *VeloVI* hanno prodotto complessivamente le ricostruzioni maggiormente concordi con la dinamica biologica attesa e attraverso l'analisi dell'intrinsic e dell'extrinsic uncertainty di *VeloVI* è stato possibile quantificare l'incertezza nelle velocità. Si è osservato che alcune regioni caratterizzate da dinamiche biologicamente poco plausibili, come quella delle popolazioni granulari nel giro dentato, presentano

anche una maggiore variabilità delle direzioni o degli stati futuri predetti. Tuttavia, una ridotta incertezza non permette di verificarne direttamente la correttezza biologica, deve essere interpretata insieme ad altre metriche e alla conoscenza del sistema biologico. Il confronto dei *Phase Portraits* tra *scVelo* e *VeloVI* ha mostrato che le due procedure possono produrre diverse ricostruzioni per lo stesso gene e che possono non descrivere adeguatamente la distribuzione dei dati soprattutto in presenza di segnali deboli o di regimi cinetici multipli.

Dai risultati emerge quindi che la sofisticazione crescente dei modelli (attraverso formulazioni bayesiane, rappresentazioni latenti condivise tra geni, reti neurali...) non è sufficiente a eliminare le principali criticità della RNA velocity. Anche gli approcci più avanzati possono infatti produrre stime instabili e backflow, suggerendo che dei limiti dipendano dalle assunzioni cinetiche condivise dai diversi metodi, dal dataset poco informativo e dalla sensibilità delle procedure di inferenza alle scelte computazionali.

Bibliografia

Alexander Aivazidis, Fani Memi, Vitalii Kleshchevnikov, et al. Cell2fate infers RNA velocity modules to improve cell fate prediction. *Nature Methods*, 22: 698–707, 2025. doi: 10.1038/s41592-025-02608-3.

Aimée Bastidas-Ponce, Sophie Tritschler, Leander Dony, Katharina Scheibner, Marta Tarquis-Medina, Ciro Salinno, Silvia Schirge, Ingo Burtscher, Anika Böttcher, Fabian J Theis, Heiko Lickert, and Mostafa Bakhti. Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development*, 146(12):dev173849, 2019. doi: 10.1242/dev.173849.

Volker Bergen, Marius Lange, Stefan Peidli, et al. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, 38: 1408–1414, 2020. doi: 10.1038/s41587-020-0591-3.

Volker Bergen, Ruslan A. Soldatov, Peter V. Kharchenko, and Fabian J. Theis. RNA velocity—current challenges and future perspectives. *Molecular Systems Biology*, 17:e10282, 2021. doi: 10.15252/msb.202110282.

Nicolas L. Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34:525–527, 2016. doi: 10.1038/nbt.3519.

Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013. doi: 10.1093/bioinformatics/bts635.

Yuheng Du, Qianhui Huang, Cedric Arisdakessian, and Lana X. Garmire. Evaluation of STAR and Kallisto on single cell RNA-Seq data alignment. *G3 Genes|Genomes|Genetics*, 10(5):1775–1783, 2020. doi: 10.1534/g3.120.401160.

Adam Gayoso, Philipp Weiler, Mohammad Lotfollahi, et al. Deep generative modeling of transcriptional dynamics for RNA velocity analysis in single cells. *Nature Methods*, 21:50–59, 2024. doi: 10.1038/s41592-023-01994-w.

Yichen Gu, David Blaauw, and Joshua D. Welch. Bayesian inference of RNA velocity from multi-lineage single-cell data. *bioRxiv*, 2022. doi: 10.1101/2022.07.08.499381.

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

Kexin Huang, Yu Zhou, Tiangang Wang, Xiao Li, Xinlong Zhao, Xi Liu, Liyu Huang, Xiaobo Zhou, and Jiajia Liu. Benchmarking algorithms for RNA velocity inference. *bioRxiv*, 2026. doi: 10.64898/2026.01.03.697314.

Gioele La Manno, Ruslan Soldatov, Amit Zeisel, et al. RNA velocity of single cells. *Nature*, 560(7719):494–498, 2018. doi: 10.1038/s41586-018-0414-6.

Atefeh Lafzi, Catia Moutinho, Simone Picelli, and Holger Heyn. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. *Nature Protocols*, 13(12):2742–2757, 2018. doi: 10.1038/s41596-018-0073-y.

Ya Luo, Jun Ren, Qian Yang, Ying Zhou, Zhiyu You, and Qiyuan Li. Benchmarking RNA velocity methods across 17 independent studies. *bioRxiv*, 2025. doi: 10.1101/2025.08.02.668272.

Páll Melsted, Vasilis Ntranos, and Lior Pachter. The barcode, UMI, set format and BUStools. *Bioinformatics*, 35(21):4472–4473, 2019. doi: 10.1093/bioinformatics/btz279.

Viktor Petukhov, Jimin Guo, Ninib Baryawno, Nicolas Severe, David T. Scadden, Maria G. Samsonova, and Peter V. Kharchenko. dropEst: pipeline for accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments. *Genome Biology*, 19:78, 2018. doi: 10.1186/s13059-018-1449-6.

Charlotte Soneson, Avi Srivastava, Rob Patro, and Michael B. Stadler. Preprocessing choices affect RNA velocity results for droplet scRNA-seq data. *PLoS Computational Biology*, 17(1):e1008585, 2021. doi: 10.1371/journal.pcbi.1008585.

Shijie C. Zheng, Genevieve Stein-O'Brien, Leandros Boukas, Loyal A. Goff, and Kasper D. Hansen. Pumping the brakes on RNA velocity by understanding and interpreting RNA velocity estimates. *Genome Biology*, 24:246, 2023. doi: 10.1186/s13059-023-03065-x.