

RESSALVA

Atendendo solicitação do(a) autor(a), o texto completo deste documento será disponibilizado somente a partir de 19/11/2026.

COMPARAÇÃO DE FERRAMENTAS CONVENCIONAIS E MODELOS DE DEEP LEARNING NA ANÁLISE DE RNA VELOCITY EM SEQUENCIAMENTO DE RNA SINGLE-CELL

MATHEUS RODRIGUES SAUDA

BOTUCATU-SP

2025

UNIVERSIDADE ESTADUAL PAULISTA

"Júlio de Mesquita Filho"

INSTITUTO DE BIOCÊNCIAS DE BOTUCATU

**COMPARAÇÃO DE FERRAMENTAS CONVENCIONAIS E
MODELOS DE DEEP LEARNING NA ANÁLISE DE RNA
VELOCITY EM SEQUENCIAMENTO DE RNA SINGLE-CELL**

Mestrando: Matheus Rodrigues Sauda

Orientador: Guilherme Targino Valente

Dissertação apresentada ao Instituto de
Biotecnologia, Câmpus de Botucatu,
UNESP, como parte dos pré-requisitos
necessários para obtenção do título de
Mestre no Programa de Pós-Graduação
em Ciências Biológicas (Genética)
IBB/Unesp

BOTUCATU-SP

2025

S255c Sauda, Matheus Rodrigues
 Comparação de ferramentas convencionais e
 modelos de Deep Learning na análise de RNA
 Velocity em sequenciamento de RNA Single-cell. /
 Matheus Rodrigues Sauda. -- Botucatu, 2025
 41 p. : il., tabs.

 Dissertação (mestrado) - Universidade Estadual
 Paulista (UNESP), Instituto de Biociências, Botucatu
 Orientador: Guilherme Targino Valente

 1. Single-cell RNA sequencing. 2. RNA Velocity. 3.
 Deep Learning. 4. Benchmarking. I. Título.

ATA DA DEFESA PÚBLICA DA DISSERTAÇÃO DE MESTRADO DE MATHEUS RODRIGUES SAUDA, DISCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS BIOLÓGICAS (GENÉTICA), DO INSTITUTO DE BIOCIÊNCIAS - CÂMPUS DE BOTUCATU.

Aos 19 de novembro de 2025, às 9h, por meio de Videoconferência, realizou-se a defesa de DISSERTAÇÃO DE MESTRADO de MATHEUS RODRIGUES SAUDA, intitulada **Comparação de ferramentas convencionais e modelos de Deep Learning na análise de RNA Velocity em sequenciamento de RNA Single-cell.** A Comissão Examinadora foi constituída pelos seguintes membros: Prof. Dr. GUILHERME TARGINO VALENTE (Orientador(a) - Participação Virtual) do(a) Laboratório de Biotecnologia Aplicada / Faculdade de Medicina de Botucatu - UNESP, Dr. MARCEL RODRIGUES FERREIRA (Participação Virtual) do(a) Unidade de Pesquisa Experimental / Faculdade de Medicina de Botucatu - Unesp, Dr. LEONARDO NAZARIO DE MORAES (Participação Virtual) do(a) Diretoria / UNESP / Câmpus de Botucatu - FCA, Após a exposição pelo mestrando e arguição pelos membros da Comissão Examinadora que participaram do ato, de forma presencial e/ou virtual, o discente recebeu o conceito final: Aprovado. Nada mais havendo, foi lavrada a presente ata, que após lida e aprovada, foi assinada pelo(a) Presidente(a) da Comissão Examinadora.

Documento assinado digitalmente
 GUILHERME TARGINO VALENTE
Data: 19/11/2025 14:22:47-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. GUILHERME TARGINO VALENTE

AGRADECIMENTOS

Agradeço à minha família pelo apoio e incentivo ao longo desses anos, em especial a Roberto, Elizângela, Juliana e Luzia.

À minha namorada e companheira, Maria, pelo apoio durante essa jornada.

À amiga e colega de laboratório Ana, e a todos os amigos fora da academia.

Ao Professor Dr. Guilherme, pela orientação e pela oportunidade de crescimento acadêmico.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela concessão da bolsa de mestrado.

Ao Programa de Pós-Graduação em Ciências Biológicas (Genética), pela oportunidade.

RESUMO

A técnica de *single-cell RNA sequencing* (scRNA-Seq) modernizou a biologia molecular ao possibilitar a análise da expressão gênica em nível de célula individual, revelando a heterogeneidade e a complexidade funcional dos tecidos. Nesse contexto, a análise de *RNA velocity* se destaca por permitir a inferência da direção e da velocidade das trajetórias celulares com base nas proporções de transcritos processados (*spliced*) e não processados (*unspliced*), fornecendo uma visão dinâmica dos estados celulares ao longo do tempo. Essa abordagem é especialmente útil em estudos de diferenciação, transição entre estados e plasticidade celular. Tradicionalmente, métodos como o scVelo utilizam modelos cinéticos baseados em taxas de transcrição e degradação específicas para cada gene. Recentemente, porém, modelos de deep learning vêm sendo aplicados a essa etapa, explorando *variational autoencoders* (VAEs) para capturar relações não lineares e complexas nos dados de scRNA-Seq. Tais modelos, como DeepVelo, VeloVI e LatentVelo, apresentam-se como alternativas promissoras às abordagens tradicionais na reconstrução de trajetórias celulares. Neste estudo, foram comparadas abordagens clássicas e baseadas em deep learning para a inferência de *RNA velocity*, utilizando dois conjuntos de dados públicos (GSE149689 de humano, e GSE203233 de camundongo). Após o pré-processamento com Cell Ranger e Scanpy, os dados foram analisados com o scVelo e com os modelos de *deep learning* mencionados. As comparações foram realizadas por meio da similaridade cosseno entre os vetores de velocidade e da análise do erro quadrático médio (MSE) para avaliar a continuidade das trajetórias celulares. Os resultados demonstraram que os métodos baseados em deep learning geraram inferências mais consistentes e biologicamente plausíveis em comparação ao modelo clássico. Em especial, o DeepVelo apresentou o menor erro de continuidade entre as abordagens de *deep learning*. Apesar das vantagens em precisão e coerência biológica, essas metodologias exigem maior poder computacional e dependem de uma quantificação precisa dos transcritos *spliced* e *unspliced*. Assim, conclui-se que os modelos baseados em VAEs representam um avanço promissor para a análise de *RNA velocity* em dados de scRNA-Seq, oferecendo novas perspectivas para o estudo de processos celulares complexos.

Palavras-chave: *Single-cell RNA sequencing, RNA Velocity, Deep Learning, Benchmarking*

ABSTRACT

Single-cell RNA sequencing (scRNA-Seq) has transformed molecular biology by enabling gene expression analysis at single-cell resolution, unveiling the heterogeneity and functional complexity of tissues. Within this framework, RNA velocity analysis stands out for its ability to infer the direction and rate of cellular trajectories based on the relative abundances of processed (spliced) and unprocessed (unspliced) transcripts, providing a dynamic view of cell-state transitions over time. This approach is particularly valuable in studies of differentiation, state transitions, and cellular plasticity. Classical methods such as scVelo employ kinetic models grounded on gene-specific transcription and degradation rates. Recently, deep learning-based approaches have been applied to this task, leveraging variational autoencoders (VAEs) to capture nonlinear and complex patterns in scRNA-Seq data. Models including DeepVelo, VeloVI, and LatentVelo have emerged as promising alternatives to traditional frameworks for reconstructing cellular trajectories. In this study, we compared classical and deep learning-based approaches for RNA velocity inference using two public datasets (GSE149689 of human, and GSE203233 of mouse). Following preprocessing with Cell Ranger and Scanpy, the data were analyzed using scVelo and the deep learning models. Comparative analyses were performed using cosine similarity between velocity vectors and mean squared error (MSE) metrics to evaluate the continuity and coherence of inferred cellular trajectories. The results demonstrated that deep learning-based methods produced more consistent and biologically plausible velocity fields compared to the classical model. In particular, DeepVelo achieved the lowest trajectory continuity error among the deep learning methods. Despite their advantages in accuracy and biological consistency, these approaches demand higher computational resources and depend on precise quantification of spliced and unspliced transcripts. Overall, VAE-based frameworks represent a promising advancement for RNA velocity analysis in scRNA-Seq data, offering new insights into complex cellular processes and dynamic transcriptional landscapes.

Keywords: Single-cell RNA Seq, RNA Velocity, Deep Learning, Benchmarking

SUMÁRIO

1. INTRODUÇÃO	9
1.1 <i>Single-Cell RNA Sequencing</i>	9
1.2 <i>RNA Velocity</i>	12
1.3 <i>Deep learning</i> e modelos presentes em <i>RNA Velocity</i>	14
2. OBJETIVO	17
2.1 Objetivo Geral	17
2.2 Objetivos Específicos	17
3. RESULTADOS	18
4. CONSIDERAÇÕES FINAIS	38
5. REFERÊNCIAS BIBLIOGRÁFICAS	39

1. INTRODUÇÃO

1.1 *Single-Cell RNA Sequencing*

A técnica de *single-cell RNA sequencing* (*scRNA-Seq*) permite a avaliação da expressão genética em um nível celular (Chen; Ning; SHI, 2019), possibilitando a observação e avaliação da heterogeneidade do tecido (Aljanahi; Danielsen; Dunbar, 2018). Além disso, a análise possibilita a identificação de novos tipos de células e a definição de subtipos celulares dentro do tecido avaliado (Aljanahi; Danielsen; Dunbar, 2018).

Comparada a outras técnicas, como o *bulk RNA-Seq*, que reflete principalmente a expressão genética média entre milhares de células (Chen; Ning; Shi, 2019) e a *qRT-PCR*, que possibilita uma análise limitada de genes com hipóteses previamente estabelecidas (Saliba *et al.*, 2014), a abordagem de *scRNA-Seq* oferece uma resolução superior em termos de heterogeneidade celular e variabilidade na expressão gênica célula à célula (Chen; Ning; Shi, 2019). Além disso, o *scRNA-Seq* permite a identificação de possíveis marcadores, como proteínas de superfície, úteis para a avaliação de tipos celulares sem marcadores conhecidos (Aljanahi; Danielsen; Dunbar, 2018).

A técnica de *scRNA-Seq* envolve, em resumo, a transcrição reversa do *RNA* em *cDNA*, seguida pela amplificação e sequenciamento do *DNA* complementar (*cDNA*) usando tecnologias de sequenciamento de nova geração (*NGS*) (Saliba *et al.*, 2014). Devido à sua natureza suscetível a ruídos e variações biológicas, como a transcrição estocástica, caracterizada pela ativação ou inativação aleatória de genes através da ligação e liberação de fatores de transcrição, a análise por *scRNA-Seq* requer múltiplas etapas para processar os dados obtidos pela técnica (Chen; Ning; Shi, 2019). Isso implica na construção de um *workflow* de *single-cell*, com o objetivo de obter resultados relevantes e reprodutíveis a partir dos dados obtidos pelo método.

A divisão por etapas no *workflow* (Fig.1) de *scRNA-Seq* possibilita uma compreensão mais aprofundada dos processamentos aos quais os dados são submetidos, além de facilitar o desenvolvimento de ferramentas para cada fase do fluxo de trabalho. O *workflow* da análise por *single-cell RNA-seq* pode ser dividido em:

- *Pre-processing and Quality Control*: Apesar da riqueza de informações que os dados de *scRNA-Seq* podem fornecer, sua complexidade e ruído são significativamente maiores em comparação aos dados de *bulk RNA-Seq* (Erfanian *et al.*, 2023). A etapa de controle de qualidade tem como objetivo remover possíveis marcações errôneas ou ausentes na identificação de populações de células por meio de códigos de células (*cell barcodes*) (Erfanian *et al.*, 2023), uma técnica que classifica células individuais com ácidos nucleicos únicos, possibilitando seu monitoramento espacial e temporal (Kebschull; Zador, 2018).
- *Normalization*: Assim como o pré-processamento, a normalização desempenha um papel crucial no tratamento de dados de *scRNA-Seq*, com o objetivo de identificar e corrigir possíveis variações entre amostras causadas por artefatos ou efeitos biológicos inesperados, visando remover possíveis vies nos dados (Erfanian *et al.*, 2023).
- *Data Correction*: Consiste na etapa de remoção de variabilidades técnicas e biológicas inesperadas, podendo ser dividido em duas etapas:
 - *Dropout*: Definidos como eventos de falsos zeros, os *dropouts* são resultado de uma baixa captura de *RNA* mensageiro (*mRNA*), levando a uma menor probabilidade de capturar genes expressos (Erfanian *et al.*, 2023). Seja devido à baixa captura de *mRNA* ou a efeitos estocásticos, os eventos de *dropout* são observados nos dados quando há moderada ou alta expressão de um gene em uma célula, mas não é observado em outra célula (Erfanian *et al.*, 2023).
 - *Batch Effect Correction*: O termo *batch effect* pode ser definido como variações causadas por fatores técnicos. Esses efeitos técnicos podem ser atribuídos a diferentes máquinas de sequenciamento, ambientes laboratoriais diversos, origens de amostra distintas e até mesmo a diferentes técnicos que realizaram a preparação da biblioteca para o *scRNA-Seq* (Erfanian *et al.*, 2023). Essas variações podem impactar a etapa de análise *downstream* e levar a conclusões

equivocadas devido à precisão alterada das análises (Erfanian *et al.*, 2023).

- *Dimensionality Reduction*: Os métodos de redução de dimensionalidade transformam os dados, originalmente com uma quantidade alta de ruídos na expressão dos genes em dimensões altas, em dados com dimensões menores. Isso possibilita análises *downstream* apropriadas e detalhadas devido à maior facilidade de observar os dados de *scRNA-Seq* e realizar inferências relevantes (Sun *et al.*, 2019).
- *Downstream Analysis*: Podem ser definidos como métodos utilizados para identificar mecanismos biológicos relevantes nos dados tratados (Erfanian *et al.*, 2023). As análises *downstream* podem ser divididas em:
 - *Clustering and cell annotation*: Esta é uma parte significativa na análise de *scRNA-Seq* que envolve a classificação de subpopulações de células em grupos com relevância biológica (Erfanian *et al.*, 2023).
 - *Cell-cell communication*: O objetivo desta etapa do *workflow* de *scRNA-Seq* é analisar a comunicação entre as células. A comunicação entre células controlada pelo complexo de ligante e receptor é relevante para observar processos biológicos como desenvolvimento, diferenciação e inflamação (Erfanian *et al.*, 2023).
 - *RNA Velocity*: Pode ser definido como a taxa de mudança na expressão dos genes em uma célula em um determinado momento, dependendo da quantidade de *mRNA* que passou pelo processo de *splicing* e quantidade que não passou pelo processo de *splicing* (Erfanian *et al.*, 2023).

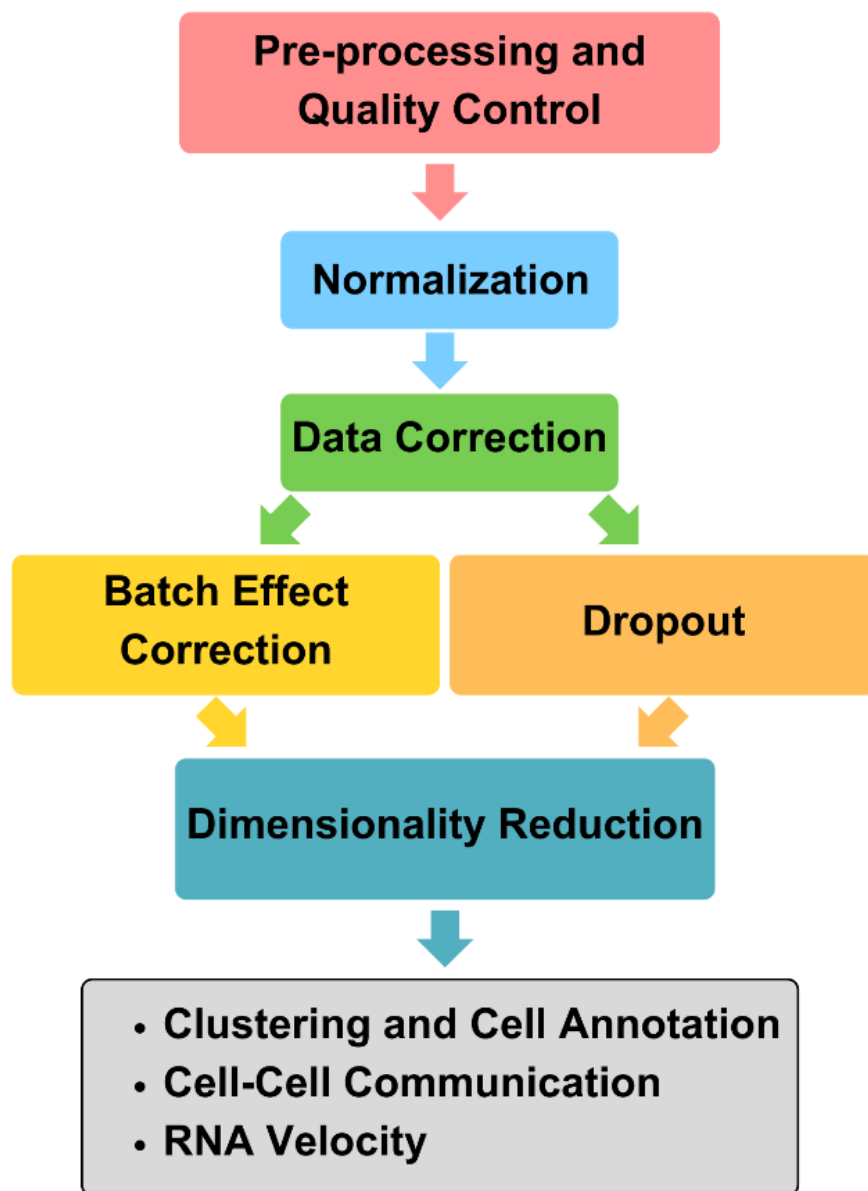


Figura 1: *Workflow* para análises de *scRNA-Seq*. **Fonte:** Feito com base no trabalho de ERFANIAN *et al.*, 2023.

5. REFERÊNCIAS BIBLIOGRÁFICAS

ALJANAHI, A. A.; DANIELSEN, M.; DUNBAR, C. E. *An introduction to the analysis of single-cell RNA-sequencing data*. **Molecular Therapy Methods & Clinical Development**, v. 10, p. 189–196, 21 set. 2018.

BERGEN, V. et al. *Generalizing RNA velocity to transient cell states through dynamical modeling*. **Nature Biotechnology**, v. 38, n. 12, p. 1408–1414, dez. 2020.

BERGEN, V. et al. *RNA velocity—current challenges and future perspectives*. **Molecular Systems Biology**, v. 17, n. 8, p. e10282, ago. 2021.

CELL RANGER - OFFICIAL 10X GENOMICS SUPPORT. Disponível em: <https://www.10xgenomics.com/support/software/cell-ranger/latest>. Acesso em: 15 jan. 2025.

CHEN, G.; NING, B.; SHI, T. *Single-cell RNA-seq technologies and related computational data analysis*. **Frontiers in Genetics**, v. 10, 2019.

CHEN, Z. et al. *DeepVelo: Single-cell transcriptomic deep velocity field learning with neural ordinary differential equations*. **Science Advances**, v. 8, n. 48, p. eabq3745, 30 nov. 2022.

CUI, H. et al. *DeepVelo: deep learning extends RNA velocity to multi-lineage systems with cell-specific kinetics*. **Genome Biology**, v. 25, n. 1, p. 27, 19 jan. 2024.

ERFANIAN, N. et al. *Deep learning applications in single-cell genomics and transcriptomics data analysis*. **Biomedicine & Pharmacotherapy**, v. 165, p. 115077, 1 set. 2023.

FARRELL, S.; MANI, M.; GOYAL, S. *Inferring single-cell transcriptomic dynamics with structured latent gene expression dynamics*. **Cell Reports Methods**, v. 3, n. 9, p. 100581, 25 set. 2023.

GAYOSO, A. et al. *Deep generative modeling of transcriptional dynamics for RNA velocity analysis in single cells*. **Nature Methods**, v. 21, n. 1, p. 50–59, jan. 2024.

KEBSCHULL, J. M.; ZADOR, A. M. *Cellular barcoding: lineage tracing, screening and beyond*. **Nature Methods**, v. 15, n. 11, p. 871–879, nov. 2018.

LA MANNO, G. et al. *RNA velocity of single cells*. *Nature*, v. 560, n. 7719, p. 494–498, ago. 2018.

LI, Q. *scTour: a deep learning architecture for robust inference and accurate prediction of cellular dynamics*. **Genome Biology**, v. 24, n. 1, p. 149, 23 jun. 2023.

OXFORD REFERENCE. *Splicing*. Disponível em: <https://www.oxfordreference.com/display/10.1093/oi/authority.20110803100524412>. Acesso em: 11 fev. 2024.

PROJECT JUPYTER. Disponível em: <https://jupyter.org>. Acesso em: 15 jan. 2025.

RIBA, A. et al. *Cell cycle gene regulation dynamics revealed by RNA velocity and deep learning*. **Nature Communications**, v. 13, n. 1, p. 2865, 23 maio 2022.

SALIBA, A.-E. et al. *Single-cell RNA-seq: advances and future challenges*. **Nucleic Acids Research**, v. 42, n. 14, p. 8845–8860, 18 ago. 2014.

SUN, S. et al. *Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis*. **Genome Biology**, v. 20, n. 1, p. 269, 10 dez. 2019.

What Is Deep Learning? | **IBM**. Disponível em: <https://www.ibm.com/think/topics/deep-learning>. Acesso em: 15 jan. 2025.

XIE, C. et al. *RNA velocity prediction via neural ordinary differential equation*. **iScience**, v. 27, n. 4, p. 109635, 19 abr. 2024.

ZHENG, S. C. et al. *Pumping the brakes on RNA velocity by understanding and interpreting RNA velocity estimates*. **Genome Biology**, v. 24, n. 1, p. 246, 26 out. 2023.