中国科学：数学　2025 年　第 55 卷　第 7 期：1～14
SCIENTIA SINICA Mathematica
综　述
《中国科学》杂志社
SCIENCE CHINA PRESS

生物高维复杂数据的动态建模、优化与深度学习

献给钱敏平教授 86 寿辰

蒋淇[1,2]，万林[1,2]*

1. 中国科学院数学与系统科学研究院, 北京 100190;
2. 中国科学院大学数学科学学院, 北京 100049
E-mail: jiangqi201@mails.ucas.ac.cn, lwan@amss.ac.cn

收稿日期: 2024-10-08; 接受日期: 2025-01-09;　* 通信作者
国家重点研发计划 (批准号: 2022YFA1004800) 资助项目

摘要　生物高维大数据的不断涌现, 其数学建模与计算对概率统计、数据科学和人工智能等领域提出了新的挑战. 本文综述了以时间序列单细胞测序为代表的快照型高维数据的动力学建模与计算的研究进展, 重点介绍了基于最优传输理论的数学模型与优化方法, 强调了数学模型与深度学习相融合在高维大数据高效计算中的关键作用, 并讨论了推断细胞 - 细胞相互作用等复杂非线性系统建模与计算方面的挑战与机遇.

关键词　快照型高维数据　动态建模　最优传输　最优控制　深度学习

MSC (2020) 主题分类　60J70, 49Q22, 37N25, 92C37

1　引言

当前, 生物高维复杂数据, 特别是单细胞和空间多组学数据的集中涌现, 为研究细胞命运决策、复杂疾病机理、人脑结构与功能等生物医学中的关键科学问题提供了重要基础 [55]. 单细胞测序技术可以在特定时间点测量单个细胞的高维分子特征 (如基因表达、蛋白、表观遗传等), 为复杂生物系统的机制与过程解析提供了更精细的数据基础. 同一个时间点获得的细胞群体数据, 通常称为快照型 (snapshot) 数据. 然而, 生物系统通常是随时间不断演化的, 仅靠单一时间点的静态快照数据不足以揭示复杂系统的动态变化. 通过在不同时间点对单细胞群体进行测序, 获得多个时间点的快照数据, 可以更深入地探索生物系统的动态演化规律.

以时间序列单细胞测序数据为代表的动态快照型数据的分析提出了新的数学与计算问题. 经典的时间序列分析是对同一个体随时间变化的轨迹进行动力学建模与分析, 使用的时间序列数据通常是对固定个体在不同的时间点的状态进行观测和记录. 然而, 单细胞测序数据的获取方法往往是破坏性的, 即每个细胞在测序后已经被破坏. 这意味着同一个细胞无法在多个时间点进行追踪观测和测量. 这就

英文引用格式: Jiang Q, Wan L. Dynamic modeling, optimization, and deep learning for high-dimensional complex biological data (in Chinese). Sci Sin Math, 2025, 55: 1–14, doi: 10.1360/SSM-2024-0308

造成了不同时刻测量的细胞数据缺少匹配关系. 因此, 需要发展新型的动态建模与计算方法, 通过学习在不同时间下的细胞群体的概率分布, 重构复杂生物系统的动态过程 (图 1). Yang 等 [65] 将对快照型的时间序列分析称为基于概率分布度量下的系综回归 (ensemble-regression), 以区别于传统时间序列分析的基于 Euclid 空间度量下的点回归 (point-regression). 此外, 生物系统通常表现出高度的非线性和随机性. 如何对时间序列单细胞高维数据进行数学建模和高效计算, 揭示复杂生物系统的演化规律, 已成为当前研究的热点问题.

以深度学习为代表的人工智能技术在生物医学大数据的解析中取得了大量应用. 然而, 纯数据驱动的深度学习方法普遍被视为 “黑箱”, 缺少良好的可解释性. 当模型遇到分布外数据 (out of distribution, OOD) 时, 其泛化能力也可能受到显著影响. 特别是在时间序列单细胞数据分析中, 常常出现分布偏移 (distribution drift), 即训练数据的概率分布与测试数据的概率分布 (即未观测时间点的概率分布) 存在显著差异. 这种差异意味着当模型需要预测未观测时间点的细胞分布时, 泛化性能可能会大幅下降, 导致不准确的预测结果. 因此, 如何发展模型与数据共同驱动的智能算法, 增强可解释性与泛化能力, 也是当前的一个重要研究方向和挑战.

时间序列单细胞数据分析的一个核心问题是如何对不同时刻的数据进行对齐与匹配. 最优传输 (optimal transport, OT) 理论提供了一种有效工具, 并逐渐被应用于处理时间序列的单细胞数据分析中. 最优传输通过最小化两个概率分布之间的运输成本, 找到从一个分布到另一个分布的最优映射. 最小运输成本可以有效量化不同概率分布之间距离. 通过将每个时间点的快照数据看作服从某个定义在基因表达空间的概率分布 (图 1(b)), 最优传输方法可以用于推断这些分布之间的最优映射, 从而在不同时间点之间建立概率关联. 该方法背后的假设是, 细胞在两个状态之间的转化遵循最小作用原理 (principle of least action) [19,22], 即系统选择的演化路径会使得作用量达到最小值. 基于这个假设, Schiebinger 等 [49] 进一步考虑了细胞的增殖与凋亡, 在相邻时刻数据之间建立了非平衡最优传输 (unbalanced OT) 模型; 并学习细胞 - 细胞的最优概率匹配, 用于估计细胞的祖先分布和后代分布. Yang 和 Uhler [64] 提出了一种可扩展的非平衡最优传输方法, 直接利用生成对抗网络 (generative adversarial network, GAN) 对映射进行参数化. 该方法能够高效处理大规模复杂数据, 降低计算成本.

然而, 直接利用传统最优传输理论寻找最优匹配的方法只能在相邻时间点之间学习静态映射, 难以捕捉细胞动力学在时间轴上的连续性特征. 近年来, 研究者开始通过动力学模型来模拟单个细胞状态随时间的演化. 例如, Tong 等 [56] 结合动态最优传输 (dynamic optimal transport, DOT) 理论和连续归一化流 (continuous normalizing flow, CNF), 利用常微分方程 (ordinary differential equation, ODE)

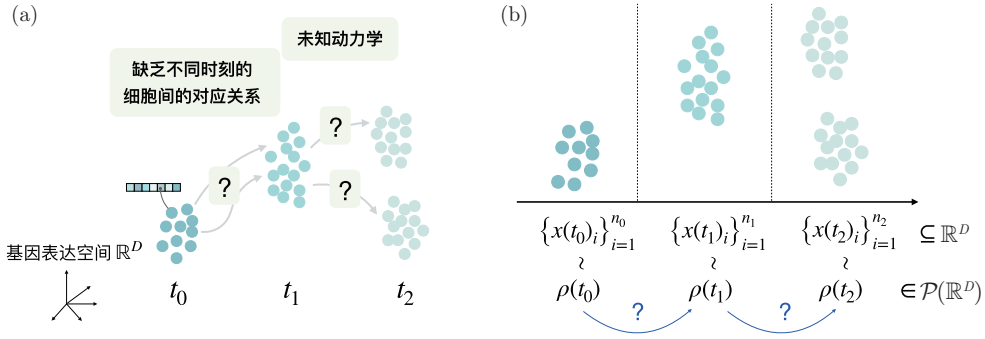

图 1 **(a) 时间序列单细胞数据带来的挑战, 其中, 每个小球表示一个细胞, 细胞状态用 D 维向量表示; (b) 将细胞状态看作是从某个定义在基因表达空间的概率分布中采样得到的, 问题转变为研究这些不同时刻的概率分布是如何在概率空间上变化的. 其中, $\mathcal{P}(\mathbb{R}^D)$ 表示定义在基因表达空间 $\mathbb{R}^D$ 上的概率空间**

描述确定性的细胞动力学. 为了进一步考虑生物过程中的随机性, Yeo 等[66] 采用由势能驱动的随机微分方程 (stochastic differential equation, SDE) 模拟细胞分化, 并设计了一个经验正则化项约束最终时刻的细胞势能. 我们指出, 这种经验正则化项仅作用于最终时刻的细胞, 可能导致模型在最终时刻出现过拟合. 为此, 我们开发了模型与数据融合的新型智能算法 (physics-informed neural SDE, PI-SDE)[28]. 通过引入一个内嵌物理信息的正则化项, PI-SDE 实现了对整个状态空间上势能函数的全局约束, 并在单细胞大数据分析中取得了更高的可解释性及可预测性.

对于大规模高维生物数据, 带有微分方程约束的优化问题通常难以通过传统的时间离散化方法有效求解, 尤其是在处理不规则采样或复杂非线性动力学时. 2018 年, Chen 等[11] 提出的神经常微分方程 (neural ordinary differential equation, Neural ODE) 提供了一种新的思路. Neural ODE 的核心思想是直接通过神经网络拟合动力学方程中的关键项, 实现对连续时间变化的模拟, 从而避免了传统离散层堆叠方式中的计算瓶颈. Neural ODE 不仅能根据需求灵活选择不同阶数的 ODE 求解器, 有效处理不规则数据, 而且运用对偶方法显著减少内存占用, 为复杂生物系统建模提供了高效且强大的工具. 自提出以来, Neural ODE 得到了广泛研究 (参见文献 [23, 47, 70]). 随后, Li 等[37] 提出的 Neural SDE 框架显著改进了对 SDE 的处理能力, 尤其是在大规模问题中的可扩展性方面. 通过改进梯度计算方法, 确保了模型在处理随机性的同时保持计算效率和精度.

现有的动态建模方法普遍基于一个简化假设, 即单个细胞的基因表达变化率仅由其自身的状态决定. 然而, 细胞命运并非在孤立环境下决定. 在复杂的多细胞系统中, 细胞内调控与细胞间通信同样重要. 研究表明, 细胞间通信在发育、免疫反应、组织修复以及疾病进展等动态生物过程中发挥关键作用 (参见文献 [1]). 由于直接在高维连续空间对大规模细胞相互作用系统建模面临巨大挑战, 现有的一些尝试主要是在粗粒化的离散细胞类型层面进行建模[29, 43]. 例如, 我们提出了基于图上的非线性 Fokker-Planck 方程的 GraphFP 算法[29], 用于重构细胞状态转变的势能景观. GraphFP 通过自由能中的非线性二次项显式引入细胞类型间的相互作用. 类似地, Park 等[43] 结合已知的受体 - 配体对信息, 通过 Bayes 框架估计了随时间变化的细胞类型间通信. Smart 和 Zilman[52] 扩展了 Hopfield 网络, 提出了一个理论模型以结合细胞内和细胞间的基因相互作用, 但该模型将基因表达简化为二进制状态, 并假设细胞间相互作用形成对称矩阵, 且仅适用于稳态条件. 未来, 如何将深度学习方法与时间序列单细胞数据相结合, 在动力学建模中考虑细胞间通信, 将成为多细胞系统建模中的重要方向.

基于本文作者的科研方向与研究内容, 本文综述当前对动态快照型复杂生物高维数据的动力学建模与计算的研究进展. 接下来的讨论将主要聚焦于基于动力学模型的方法, 并将其大致分为两类. 第一类方法假设细胞的基因表达变化率仅由其自身的状态决定, 对单细胞系统进行动力学建模. 这类方法常与神经网络相结合, 已在多个应用中取得了显著成果. 第二类方法则在建模中考虑了细胞间通信, 研究多细胞系统的动态过程. 这类方法主要是在细胞类型层面进行建模以实现高效求解. 目前, 关于细胞间通信的建模仍处于早期阶段, 但这一方向在深入理解多细胞系统动态行为中具有重要潜力. 本文介绍的方法不仅适用于时间序列单细胞数据的分析, 还可为其他涉及时间序列快照数据的领域提供参考. 例如, 在社会学中, 研究者通过观察和记录不同时间点上人群的集体行为来分析拥挤现象或疾病传播的动力学特征; 在生态行为学中, 这些方法可用于研究鱼群或鸟群的集体迁徙和群体决策.

2 最优传输

为了解决大规模高维生物快照数据的建模问题, 本节将介绍有关最优传输的相关理论—OT 和

DOT. 这些理论在处理时间序列单细胞数据的动态建模过程中至关重要. 通过理解这些基础概念, 我们能够更好地理解后续讨论的具体动力学模型.

2.1 OT 理论

OT 最初由法国数学家 Monge[39] 在 18 世纪提出, 旨在解决如何在最小化运输成本的前提下将物体从一个位置移动到另一个位置的问题. 20 世纪中期, Kantorovich[33] 通过引入线性规划的思想, 为这一问题提供了更为严格的数学基础, 最终发展成了现代的最优传输理论. 该理论提供了一种有效量化不同分布之间距离的方式, 为理解和分析分布间的变换提供了一种有效的数学工具.

最优传输的核心在于找到一种最优的传输方法, 将一种概率分布转变为另一种, 从而最小化传输成本. 参考文献 [59], 这里给出 Monge-Kantorovich 问题的具体数学形式. 给定两个分布 μ 和 ν, 以及一个定义在状态空间的成本函数 $c(x,y)$, 最优传输问题通常可以被表述为寻找一种传输方案 π, 使得从分布 μ 转移到分布 ν 的成本最小化, 即分布 μ 和 ν 的最优传输距离定义为

$$\min_{\pi\in\Pi(\mu,\nu)}\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi(x,y), \tag{2.1}$$

其中, $\Pi(\mu,\nu)=\{\pi \mid \int_{\mathcal{Y}} d\pi(x,y)=\mu(x), \int_{\mathcal{X}} d\pi(x,y)=\nu(y)\}$ 表示所有以 μ 和 ν 为边际分布的联合分布. $c(x,y)$ 是传输成本, 其中 x 和 y 分别是两个分布中的点, 常见的选择是 Euclid 距离的平方 $c(x,y)=\|x-y\|^2$, 对应的最优传输距离被称作 Wasserstein 距离.

传统最优传输问题的求解在过去几十年中取得了显著进展. 早期的静态最优传输以线性规划的形式呈现, 但计算成本相对较高, 为 $\mathcal{O}(n^3)$, 其中, n 为样本个数. 这种高昂的计算代价促使了多种快速近似算法的开发. 例如, Cuturi[16] 利用熵正则化将经典的最优传输问题转化为更易解的凸优化问题, 提出了基于矩阵迭代优化的高效求解算法—Sinkhorn 算法[51]. Muzellec 和 Cuturi[40] 进一步优化这一框架, 提出了一种在子空间中的求解方法, 该方法通过将原始高维问题映射到较低维的子空间中, 从而有效降低了计算复杂度.

在时间序列单细胞数据建模中, 可以将每个观测时间点的群体细胞状态视作一个高维点云 (对应一个概率分布), 其中每个点代表单个细胞在基因表达上的信息. 最优传输不仅可以用于研究不同时间点间细胞状态的最优映射[49,64], 还能用于评估预测的细胞状态分布与实际观测到的分布之间的差距, 为模型性能提供了直观的度量[20,28,66,67].

2.2 DOT 理论

2000 年, Benamou 和 Brenier[2] 定义并探索了最优传输的动态版本—DOT. 这一理论进一步扩展了经典最优传输问题, 为建模提供了更丰富的工具和视角. DOT 的核心思想在于引入时间维数, 这使我们能够捕捉概率分布在时间上的连续演化, 并研究不同分布之间的最优传输路径.

在这一框架下, Benamou 和 Brenier[2] 提出了一种与流体动力学相关的视角, 将最优传输距离与动力学方程联系起来:

$$\min_{f}\int_0^1\int_{\mathbb{R}^D}\frac{1}{2}\|f(x,t)\|^2\rho(x,t)dxdt \tag{2.2}$$

$$\text{s.t. }\frac{\partial\rho(x,t)}{\partial t}+\nabla\cdot(f(x,t)\rho(x,t))=0, \tag{2.3}$$

$$\rho(x,0)=\mu(x),\quad \rho(x,1)=\nu(x), \tag{2.4}$$

其中, $\rho(x,t):\mathbb{R}^D\times[0,1]\to\mathbb{R}$ 表示 t 时刻的概率密度, 连续光滑函数 $f(x,t):\mathbb{R}^D\times[0,1]\to\mathbb{R}^D$ 表示定义在状态空间的时变速度场, 方程 (2.3) 是流体力学中的连续性方程, 描述了概率密度随时间变化的质量守恒性. 动态优化目标函数 (2.2) 通过寻找满足方程 (2.3) 的 f, 使得总动力学能量最小 (minimum total kinetic energy).

求解这样带偏微分方程 (partial differential equation, PDE) 约束的优化问题, 我们就找到了最优速度场 $f(x,t)$, 确保概率密度 $\rho(x,t)$ 以 μ 作为初始分布, 在最终时间点达到 ν. 并且, 分布演化过程中的累积运输成本最小, 即总动力学能量最小. 特别地, 当运输成本为 Euclid 距离的平方时, Benamou-Brenier 公式表明, DOT 定义的 L_2-Wasserstein 距离 (最小传输成本) 与经典最优传输距离等价 [2].

早期对 DOT 问题的求解通常基于对时间和空间的离散化. 通过将连续时间和空间划分为有限的网格, 可以将原本带有 PDE 约束的连续优化问题转化为离散优化问题. 这种离散化的处理方法有效构造出凸优化问题, 让问题更易于求解, 尤其是在低维场景下. 然而, 随着维数的增加, 网格数呈指数增长, 导致计算成本显著增加 (详见文献 [46, 第七章]). 鉴于生物数据的高维性和大规模特征, 离散化方法在处理这些高维数据时难以适用. 为了解决高维数据动态建模中的计算瓶颈问题, 近年来出现了很多利用神经网络拟合高维速度场的研究 [25,56,68]. 通过神经网络学习连续函数, 可以高效地处理复杂的动力学结构, 避免离散化带来的维数灾难. 这一方法为求解 DOT 提供了一条全新的解决思路, 并在处理高维生物数据时展现出显著优势.

3 针对时间序列单细胞数据的动力学模型

早期对时间序列生物数据的分析主要集中在伪时间分析上. 这些方法往往忽略了时间信息, 假设系统处于稳定状态, 通过降维技术构建图或树结构以推断细胞轨迹 [13,53,57,63]. 近年来, 越来越多的研究开始采用微分方程来建模细胞状态随时间的演化. 通过引入时间变量, 可以模拟细胞群体随时间演化的轨迹, 推断其在不同时刻的状态, 揭示基因表达变化背后的潜在调控机制. 本节简要综述基于单细胞系统和多细胞系统的动力学建模方法.

3.1 问题描述

首先, 假设在 $T+1$ 个时间点采集的时间序列单细胞数据如下所示:

$$(t_0,X_{t_0}),(t_1,X_{t_1}),\ldots,(t_T,X_{t_T}), \tag{3.1}$$

其中, $X_{t_l}=\{x_{t_l,i}\}_{i=1}^{N^l}\in\mathbb{R}^{N^l\times D}$ 是一组在时间点 t_l $(l=0,1,\ldots,T)$ 测量的 N^l 个细胞的 D 维基因表达向量. 在时间序列单细胞数据的动态建模中, 通常将每个观测时间点的群体细胞状态看作是从某个概率分布采样得到的. 对每个观测时间点 t_l, 将 X_{t_l} 服从的经验分布表示为 $\hat{\rho}_{t_l}$.

这里讨论的对时间序列单细胞数据的动态建模主要聚焦于如何在概率空间 $\mathcal{P}(\mathbb{R}^D)$ 上找到一条“路径”, 依次连接观测时间点的经验分布 $\{\hat{\rho}_{t_0},\hat{\rho}_{t_1},\ldots,\hat{\rho}_{t_T}\}$. 从数学角度说, 就是找到一个随机过程 $\{x_t\sim\rho_t\mid t_0\leqslant t\leqslant t_T\}$, 使其在观测时刻 $\{t_0,t_1,\ldots,t_T\}$ 和经验分布相同:

$$\rho_{t_l}=\hat{\rho}_{t_l},\quad l=0,1,\ldots,T. \tag{3.2}$$

3.2 基于单细胞系统的动力学方程

在实际应用中, 研究者使用不同的动力学方程刻画随机过程随时间的动态演化. 本小节主要讨论各种基于单细胞系统的动力学方法 (图 2). 早期的工作通常采用常微分方程描述细胞状态的演化轨迹 [56]. 为了进一步学习到更符合生物意义的速度场, Tong 等 [56] 指出方程 (3.2) 只考虑对观测时间点的分布拟合, 并未对整体动力学作任何假设. 为此, 他们基于最小作用原理 [19,22], 在损失函数中加入向量场的 L_2- 范数积分作为路径能量. 整个优化问题对应于多时间边际分布约束的 DOT 问题. 对于这类具有严格的边界约束条件的微分方程, 直接求解具有挑战性. 通常的处理方式是通过放松这些边值约束, 转而最小化模型预测分布与观测边际分布之间的某种距离度量, 实现逐渐逼近. Tong 等采用 Kullback-Leibler (KL) 散度作为度量, 最终求解的问题为

$$\min_{f}\ (t_T - t_0)\int_{t_0}^{t_T}\int_{\mathbb{R}^D}\|f(x_t,t)\|^2\rho(x,t)dxdt + \lambda\sum_{l=1}^{T}\mathrm{D}_{\mathrm{KL}}(\rho_{t_l},\hat{\rho}_{t_l}) \tag{3.3}$$

$$\text{s.t.}\quad dx_t = f(x_t,t)dt, \tag{3.4}$$

$$x_{t_0}\sim\hat{\rho}_{t_0},\quad \rho_{t_l} = \mathrm{Law}(x_{t_l}), \tag{3.5}$$

其中, f 是向量场, 代表基因表达的变化方向和大小; $\mathrm{D}_{\mathrm{KL}}(P\|Q) = \int P(x)\log\frac{P(x)}{Q(x)}dx$ 表示 KL 散度, 度量分布 P 和 Q 之间的距离; $\lambda > 0$ 是超参数; $\mathrm{Law}(x)$ 表示随机变量 x 的概率分布函数; $\hat{\rho}_{t_l}$ 表示在观测时刻 t_l 的经验分布.

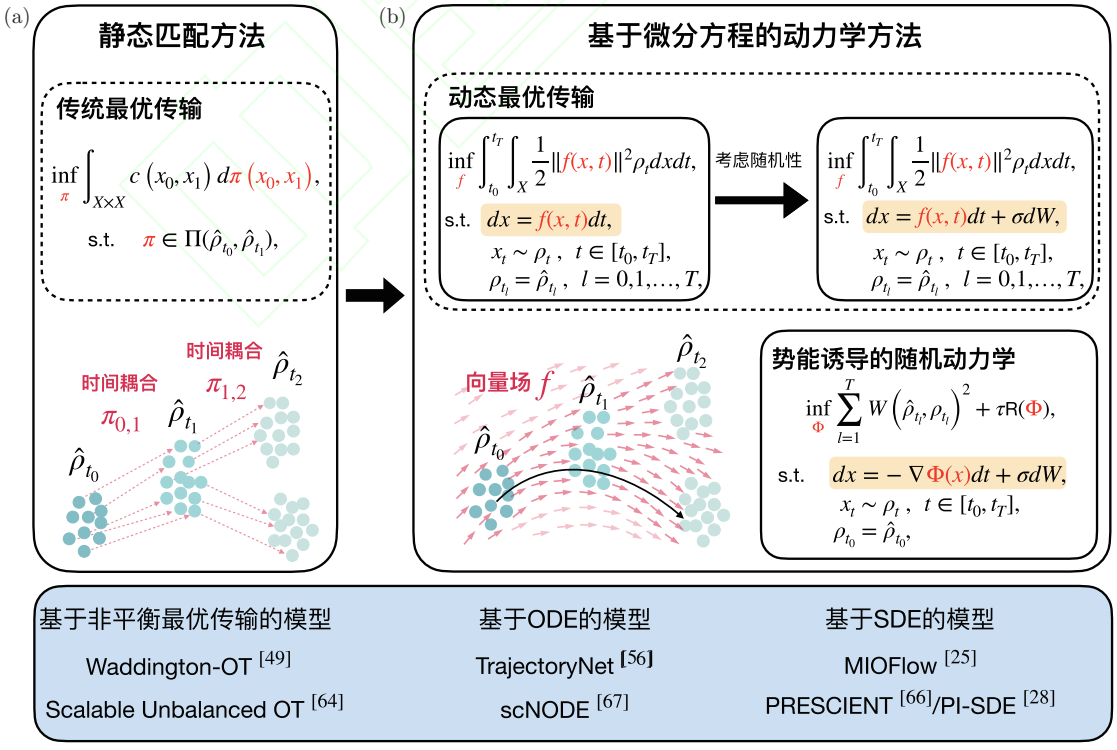

图 2 时间序列单细胞数据建模方法. (a) 静态匹配方法; (b) 基于微分方程的动力学方法. 图中展示了核心数学模型及其在动态演化建模中的关键特点, 提供了对相关方法的整体视图

为了考虑随机性, 越来越多的研究通过随机微分方程对复杂生物过程进行随机动力学建模 (参见文献 [25, 28, 66, 68]). 其中, PRESCIENT (potential energy underlying single cell gradIENTs) [66] 和 PI-SDE [28] 进一步假设随机微分方程中的漂移项 - 向量场 f 由某个势能函数 $\Phi : \mathbb{R}^D \times \mathbb{R} \to \mathbb{R}$ 诱导:

$$f(x_t, t) = -\nabla_x(\Phi(x_t, t)), \tag{3.6}$$

其中, 势能函数 Φ 驱动系统的整体动力学, 并被用于刻画英国生物学家 Waddington [60] 等提出的表观遗传景观 (Waddington's epigenetic landscape) [28, 66]. 根据 Waddington 景观, 细胞处于较高势能的状态时, 通常会向低势能状态转变. 换句话说, 细胞在最终时刻相较于早期时刻更加稳定, 所以应具有较低的势能. 因此, PRESCIENT [66] 设计了一个经验正则化项约束处在最终时刻的细胞的势能, 并采用 Wasserstein 距离度量预测分布和真实分布在观测时间的差异, 最终的动态优化问题为

$$\min_{\Phi} \sum_{l=1}^{T} \mathcal{W}(\rho_{t_l}, \hat{\rho}_{t_l}) + \tau \sum_{i=1}^{N_T} \frac{\Phi(x_{t_T,i})}{\sigma^2} \tag{3.7}$$

$$\text{s.t.} \quad dx_t = -\nabla_x(\Phi(x_t))dt + \sigma dW_t, \tag{3.8}$$

$$x_{t_0} \sim \hat{\rho}_{t_0}, \quad \rho_{t_l} = \text{Law}(x_{t_l}), \tag{3.9}$$

其中, $\mathcal{W}(P, Q)$ 表示分布 P 和 Q 的 Wasserstein 距离, Φ 刻画与细胞状态相关的势能函数, 超参数 $\sigma > 0$ 表示数值型扩散系数, W_t 表示 Brown 运动, $\tau > 0$ 是预先设定的超参数, $\text{Law}(x)$ 表示随机变量 x 的概率分布函数, $\hat{\rho}_{t_l}$ 表示在观测时刻 t_l 的经验分布.

我们指出这种经验正则化项仅作用于最终时刻的细胞, 未能对整个状态空间进行全局约束, 可能导致模型在最终时刻出现过拟合. 为了解决这一问题, 我们开发了模型和数据共同驱动的机器学习方法 PI-SDE [28]. 具体而言, PI-SDE 假设系统演化遵循最小作用原理. 同时, 根据能量景观理论和大偏差原理, 当扩散项系数足够小时, 势能函数应满足 Hamilton-Jacobi (HJ) 方程 [6, 17, 21, 27, 38, 61, 62]. PI-SDE 采用 Ruthotto 等 [48] 和 Onken 等 [41] 推导出的 HJ 方程的简化形式:

$$\partial_t \Phi(x, t) = \frac{1}{2} \|\nabla_x \Phi(x, t)\|^2. \tag{3.10}$$

然而, 直接求解包含 HJ 方程约束的优化问题极具挑战性, 尤其是随着状态空间维数的增加, 数值求解的计算复杂度呈指数增长. 为此, PI-SDE 采用一个基于物理信息的正则化项, 用于惩罚势能函数偏离 (不满足) HJ 方程的情形. 最终的优化问题为

$$\min_{\Phi,\sigma} \sum_{l=1}^{T} \mathcal{W}(\rho_{t_l}, \hat{\rho}_{t_l}) + \tau \int_{t_0}^{t_T} \int_{\mathbb{R}^D} \left| \partial_t \Phi(x, t) - \frac{1}{2} \|\nabla_x \Phi(x, t)\|^2 \right| d\rho_t(x) dt \tag{3.11}$$

$$\text{s.t.} \quad dx_t = -\nabla_x(\Phi(x_t, t))dt + \sigma(x_t, t)dW_t, \tag{3.12}$$

$$x_{t_0} \sim \hat{\rho}_{t_0}, \quad \rho_{t_l} = \text{Law}(x_{t_l}), \tag{3.13}$$

其中, 势能函数 Φ 和扩散系数 σ 都采用与细胞状态和时间都相关的函数, $\text{Law}(x)$ 表示随机变量 x 的概率分布函数, $\hat{\rho}_{t_l}$ 表示在观测时刻 t_l 的经验分布. 在实际应用中, 直接积分损失函数中的物理信息正则化项的计算量非常大, 因为在每个时间步都需要对整个空间进行积分. 为此, PI-SDE 采用路径积分 (path integration) 的方法, 即只需对初始点的状态进行采样, 再沿其轨迹计算积分即可. 通过这种方式, 我们可以避免对全局空间的每一时间步采样, 显著提高计算效率. 在单细胞大数据的分析中, PI-SDE 取得了更高的可解释性及可预测性, 并且显著提升模型训练的稳定性.

对于时间序列单细胞数据, 尽管势能诱导的随机动力学模型在捕捉细胞状态演化和转移路径方面表现出色, 但在实际应用中, 需要考虑非平衡驱动力的影响. 例如, 基因调控网络的驱动力可能受到环境变化或信号通路的动态调节. 这些因素通常无法用单一的势函数梯度表示. 为了解决这一问题, Kang 和 Li [32] 提出了一种基于非平衡态势函数的随机动力学建模方法. 该方法通过结合能量景观理论和降维技术, 能够从高维系统中提取关键的非平衡态动力学特征, 并将其投影到低维空间进行分析. 他们通过计算动力学转移路径, 捕捉到由旋流项引起的路径偏离, 明确揭示了系统中的非平衡性. 这种方法在多稳态系统的转移路径、稳定状态分布和全局动力学建模方面具有显著优势.

为了进一步模拟转录的爆发性 (bursty)、非连续的状态转变, 跳跃扩散过程逐渐被应用到生物系统的建模中 [12, 20, 24]:

$$dx_t = f(x_t, t)dt + \sigma(x_t, t)dW_t + c(x_t, t)dJ_t, \tag{3.14}$$

其中, f 表示漂移项, 代表基因表达的变化方向和大小; σ 表示扩散系数, 描述细胞运动中的随机性, 依赖于 Brown 运动 W_t; $c(x_t, t)$ 则是跳跃幅度函数, 对应复合 Poisson 过程 J_t. 跳跃扩散过程不仅包含连续的 Brown 运动, 还引入了复合 Poisson 过程来模拟突发的大幅度跃迁. 这些跃迁有助于描述细胞状态在不同吸引子之间的转换, 提供了一种更为完整的方式来捕捉复杂生物系统中的多样性和不确定性.

除了上述提到的模型外, 还有许多其他方法致力于更准确的重构细胞轨迹. 例如, Zhang 等 [68] 指出, 现有的基于 SDE 的模型虽然严格满足初始分布的约束, 但在处理后续所有观测时间点的约束时往往采用了松弛策略. 这种处理方式容易引发误差累积 (aggregation of errors), 即当前观测时间点的预测精度依赖于前一时间点的准确性. 为此, 他们引入了 Vargas 等 [58] 提出的前向 - 后向随机微分方程 (forward-backward SDE, FBSDE) 框架, 确保模型不仅满足初始时刻的分布约束, 也严格符合终止时刻的分布要求, 从而有效减小了误差累积带来的影响. 然而, FBSDE 需要对每个时间步估计模拟的分布, 使得模型在高维数据的计算精度和复杂度面临挑战. Sha 等 [50] 基于 Wasserstein-Fisher-Rao (WFR) 距离 [14], 提出了一个动态非平衡最优传输模型. 该方法完全从数据出发, 在动态建模中考虑细胞的增殖凋亡, 最终可以用于推断基因调控网络以及与细胞生长相关的关键基因. 此外, Huguet 等 [25] 设计了一个测地线自编码器 (geodesic autoencoder), 确保学习到的概率流在隐空间能够遵循原始流形的几何结构, 即在隐空间的 L_2 范数与原流形上的测地线距离相等. Zhang 等 [67] 指出, 如果未观测时间点的分布与观测时间点的分布存在显著差异, 通过观测时间点训练得到的固定低维表示可能无法很好地推广至未观测时间点. 为此, 他们引入了一个动态正则项, 将细胞的动态演化纳入隐空间, 从而生成一个信息丰富且对分布偏移具有鲁棒性的隐空间表示. 还有一类方法将时间序列单细胞数据的建模与 Schrödinger 桥 (Schrödinger bridge) 问题联系在一起. Bunne 等 [4] 通过 Gauss Schrödinger 桥研究了单细胞时间序列数据的随机动力学建模问题. Zhang 等 [69] 将 Schrödinger 桥问题与不平衡最优传输结合, 对随机系统的时间演化过程进行了重构和求解.

3.3 基于多细胞系统的动力学模型

上述的动态建模方法都基于一个简化假设, 即单个细胞的基因表达变化率仅依赖于其自身的状态. 然而, 在复杂的多细胞系统中, 细胞命运不仅取决于内部基因调控, 还受到周围细胞间通信的显著影响. 已有许多方法致力于研究如何从静态基因表达数据中推断细胞间的相互作用 [7, 18, 30]. 它们大多依赖于现有受体 - 配体数据库, 通过已知蛋白复合物的表达水平推断细胞间通信. 但如何从时间序列数据中推断细胞间相互作用仍处于起步阶段.

考虑细胞间相互作用的方程可以用非线性偏微分方程 (非线性 Fokker-Planck 方程) 描述 [9]:

$$\frac{\partial \rho(x,t)}{\partial t} = \nabla \cdot (\rho \nabla \Phi) + \nabla \cdot (\rho \nabla W * \rho) + \beta \Delta \rho, \tag{3.15}$$

其中, $W : \mathbb{R}^D \to \mathbb{R}$ 表示相互作用势能; $*$ 表示卷积运算; $\beta > 0$ 是扩散系数; Δ 是 Laplace 算子, 作用在密度上表示密度的扩散行为. 相较于传统的 (线性) Fokker-Planck 方程, 方程 (3.15) 加入了 $\nabla\cdot(\rho\nabla W*\rho)$ 项, 表示由于细胞间相互作用引起的概率流的变化. 但是, 非线性项的存在使得解析解难以获得, 数值求解时也容易导致不稳定性. 并且, 随着维数的增加, 数据空间的体积呈指数增长, 导致计算复杂度大幅上升.

近年来的研究指出, Fokker-Planck 方程可以被视为定义在 Wasserstein 流形上的梯度流 (gradient flow), 即对应于某个定义在概率分布的 Riemann 流形上自由能函数梯度的演化过程 [31, 42]. 针对 Wasserstein 梯度流, Jordan-Kinderlehrer-Otto (JKO) 框架提供了一种高效求解的方法 [3, 31, 45]. JKO 框架的核心思想是, 在每一个时间步长上, 将演化问题转化为一个变分问题, 寻找距离上一步最优解最近且能量最低的状态. 这种离散化的时间推导方法, 能够有效处理带有复杂动力学约束的优化问题. Bunne 等 [5] 通过 JKO 框架及其近端算法 (proximal algorithm), 将复杂的 Wasserstein 梯度流问题分解为一系列易于求解的子问题, 实现高效求解.

我们采用的策略是在细胞类型层面上建模动力学. 具体而言, 首先对细胞进行聚类, 并基于方程 (3.15) 的离散形式—图上的非线性 Fokker-Planck 方程 [15, 36], 我们开发了 GraphFP 算法 [29], 将 PDE 约束的优化问题转变成了 ODE 约束的优化问题, 大大降低了计算量. 该方法通过自由能刻画系统在细胞类型上的势能景观, 并通过自由能中的非线性二次项显式描述细胞间通信对其动力学的影响. 具体地, 假设细胞被分成 n 个细胞类型, 构建一个全连接的加权有向图 $G = (V, E, W)$, 其中顶点集 $V = \{1, 2, \ldots, n\}$ 代表 n 个细胞类型, 每条有向边 $\{i, j\}$ 的权重 W_{ij} 表示细胞类型 i 对细胞类型 j 的通信强度 (权重为 0 则表示没有通信). 基于此, 考虑一个定义在图 G 顶点集的离散概率空间, 称作概率单纯形:

$$\mathcal{P}(G) = \left\{ p = (p_i)_{i=1}^n \,\middle|\, \sum_{i=1}^n p_i = 1, p_i \geqslant 0 \right\}. \tag{3.16}$$

每个时刻的细胞群体可以根据其细胞类型标签表示为 $\mathcal{P}(G)$ 中的一个离散经验分布. 将每个观测时间点 t_l 的细胞群体服从的离散经验分布表示为 $\hat{p}_{t_l}$. 问题转变为如何在离散概率空间中建立细胞类型频率随时间变化的动态模型.

类似于 TrajectoryNet 的框架, 我们也最小化路径能量, 并采用 KL 散度对边际分布约束进行放松, 最终求解的问题为

$$\min_{\{\Phi, W\}} \int_{t_0}^{t_T} \frac{1}{2} \langle \nabla\Phi + \nabla(Wp) + \beta \log p, \nabla\Phi + \nabla(Wp) + \beta \log p \rangle_G dt + \sum_{l=1}^{T} \lambda_l \mathrm{D}_{\mathrm{KL}}(p(t_l) \| \hat{p}_{t_l}) \tag{3.17}$$

$$\text{s.t.} \quad \frac{dp(t)}{dt} = \nabla_G \cdot (p(t) \nabla\Phi) + \nabla_G \cdot (p(t) \nabla(Wp(t))) + \beta \Delta p(t), \tag{3.18}$$

$$dp(t_0) = \hat{p}_{t_0}, \tag{3.19}$$

其中, $\Phi \in \mathbb{R}^n$ 表示细胞类型上的线性势能, $W \in \mathbb{R}^{n\times n}$ 表示细胞类型间的相互作用势能, $\beta > 0$ 是扩散系数, $\mathrm{D}_{\mathrm{KL}}(p\|q) = \sum_i p_i \log \frac{p_i}{q_i}$ 表示离散分布 p 与 q 之间的 KL 散度, $\langle \cdot, \cdot \rangle_G$ 表示定义在图上的内积, ∇_G 表示定义在图上的散度算子 (具体定义详见文献 [15, 36]).

整个问题相当于构建了一个放松边际约束的图上的非线性最优传输问题. 在数学上, 也可以看作为一个最优控制问题. 为了评估并估计出最优参数 $\{\Phi^*_{n\times 1}, W^*_{n\times n}\}$, 我们采用 Pontryagin 极大值原理 [8]. 最终, 估计出的参数展现出高的生物可解释性, 其中, $\Phi^*_{n\times 1}$ 反映细胞的干细胞特征, $W^*_{n\times n}$ 刻画细胞间相互作用强度.

此外, Park 等 [43] 使用 Gauss 过程回归模型对细胞类型上的频率随时间的变化进行建模. 该方法结合已知的受体 - 配体对信息, 通过 Bayes 方法将细胞类型间的相互作用建模为概率问题, 捕捉随时间变化的相互作用. 它考虑的是一个线性 Gauss 回归模型, 不是在微分方程框架下的建模, 因此这里不再具体介绍. 以上两种方法仅限于在粗粒化的离散细胞类型层面进行建模. 未来, 如何将机器学习方法与时间序列单细胞数据相结合, 在动力学建模中考虑细胞间通信, 将成为多细胞系统建模中的重要方向.

4 基于深度学习的高效求解算法

在成功建模复杂生物过程后, 特别是未对连续高维空间进行简化处理的情形, 如何高效求解带有动力学方程约束的优化问题成为下一个关键挑战. 鉴于生物数据的高维数和大规模的特性, 传统数值求解方法在处理这些问题时面临计算瓶颈. 因此, 越来越多的研究方法借助深度学习, 利用神经网络来参数化动力学方程中的关键项 (如漂移项、扩散项), 以实现可计算性和高效性.

传统的深度学习模型, 如残差网络 (ResNet)、递归神经网络 (recurrent neural network, RNN) 和归一化流 (normalizing flow) 等, 通过一系列离散的网络层对 $x(t)$ 的演化进行数值近似求解, 其形式通常为 $x_{t+1} = x_t + f(x_t; \theta_t)$. 这种方法可以看作是采用了 Euler 离散化方法的近似 ODE. 随着网络层数的增加, 精度得到提升, 计算代价和内存需求也会显著增加.

CNF 是一种处理高维时间序列建模的有效工具. 例如, TrajectoryNet [56] 利用 CNF 学习随时间变化的概率密度. 传统的归一化流 (normalizing flow, NF) 通过一系列参数化的可逆变换, 将一个简单的初始分布逐步转换为复杂的目标分布. CNF 进一步将这种框架扩展到连续时间, 其变换过程由神经网络参数化的 ODE 来描述, 实现了从离散到连续的转变. 然而, CNF 在计算过程中存在一定的限制. 因为在计算真实观测分布与预测分布之间的负对数似然损失 $\sum_{l=0}^{T} \mathbb{E}_{x\sim\hat{\rho}_{t_l}}\{-\log \rho_{t_l}(x)\}$ (等价于使用 KL 散度度量观测分布与预测分布的距离) 时, CNF 利用了以下等式:

$$\log p_z(z) = \log \rho_{t_l}(x(z, t_l)) + \int_{t_{\text{init}}}^{t_l} \operatorname{tr}(\nabla f) dt, \quad t_{\text{init}} < t_0, \quad z \sim \mathcal{N}(0, 1), \tag{4.1}$$

其中, $x(z, t_l)$ 表示当 z 作为在 t_{init} 的初始状态时, 细胞在时间点 t_l 的状态. 由此可见, 模型需要计算 Jacobi 矩阵的迹 (方程 (4.1) 中的 $\operatorname{tr}(\nabla f)$), 这项计算的复杂度为 $\mathcal{O}(d^2)$, 对高维数据而言代价较大. 并且, CNF 要求从简单分布 (如 Gauss 分布) 开始, 否则方程 (4.1) 中的 $\log p_z(z)$ 项难以计算. 这种要求限制了 CNF 对真实世界中复杂分布的拟合能力, 因为所有预测的分布都必须强制从 Gauss 分布开始, 无法灵活适应数据的分布特点.

基于这种限制, 现在越来越多的方法采用 Wasserstein 距离度量真实分布和预测分布在观测时间点的差异, 并基于神经微分方程 (neural differential equation, NDE) 进行高效求解, 如求解常微分方程的 Neural ODE [11] 和求解随机微分方程的 Neural SDE [37]. Neural ODE 的一个重要贡献是能够高效处理不规则时间序列数据和高维动态系统, 尤其是在面对数据稀疏或采样不均的情况下. 使用初始状

态 $x(t_0)$ 的情况下, 可以通过积分神经网络拟合的速度场来预测任意时间点 t 的状态:

$$x(t) = x(t_0) + \int_{t_0}^{t} f(x(\tau), \tau; \theta) d\tau. \tag{4.2}$$

这种方式不再依赖于固定的时间步长, 并且能够提供高精度的函数近似. 换而言之, Neural ODE 允许根据具体需求, 灵活选择不同阶数的 ODE 求解器, 如 Euler 法和 Runge-Kutta 方法等, 从而灵活调整时间步长. 这不仅提高了计算效率, 还能够在精度和计算资源消耗之间实现更好的平衡. 另外, Neural ODE 运用对偶方法, 在反向传播过程中仅需要存储必要的状态信息, 而非所有时间步的中间状态. 这一操作显著减少了内存占用, 避免了传统神经网络在长时间序列或高维数下常见的内存瓶颈问题, 使得 Neural ODE 能够高效处理大规模数据. 同时, NDE 不需要从特定的 Gauss 分布开始, 这使得它在处理真实世界中复杂分布时更加灵活, 避免了 CNF 对 Gauss 分布的依赖.

Neural ODE 结合了深度学习的高容量函数逼近能力与数学模型的理论优势, 既能有效处理不规则数据, 又兼顾内存效率和计算成本, 自提出以来就得到了广泛研究, 并展现出强大的应用潜力. Grathwohl 等 [23] 基于 Neural ODE 框架, 提出了一种自由形式连续动力学生成模型, 该模型实现了可扩展、可逆的生成过程. Rubanova 等 [47] 提出的 Latent ODE 模型通过在隐空间中建模时间序列, 能够灵活处理不规则采样的时间序列数据. Zhong 等 [70] 通过引入可微接触模型, 提升了在不连续接触动力学中的模拟能力. Chen 和 Li [10] 将深度神经网络与部分自洽平均场近似 (partial self-consistent mean field approximation, PSCA) 相结合, 提出了一种数据驱动方法, 用于从时间序列数据中学习系统动力学方程, 推断基因调控网络, 并量化高维基因网络的能量景观. 这些研究不仅推动了 Neural ODE 在物理、生物、金融等领域的应用, 也在理论上扩展了深度学习对动态系统建模与预测的能力.

5 结论和展望

随着单细胞测序技术的快速发展, 动态建模在生物学中的应用逐渐深入, 特别是在理解复杂生物系统的演化机制方面取得了显著的进展. 基于最优传输和微分方程的建模方法, 结合深度学习技术, 已经成为处理和分析高维快照型数据的重要工具. 此外, 生物学中的时间序列数据, 尤其是破坏性的单细胞快照数据, 为数理模型和深度学习算法提出了全新的问题与挑战. 如何有效处理大规模数据集的计算瓶颈, 如何从不完整或噪声较大的数据中提取有意义的信息, 仍是当前需要解决的重要问题.

目前, 大多数方法主要考虑细胞自身的状态变化, 忽视了细胞间的复杂通信. 在机器学习领域, 这类问题被归类为多智能体系统的交互学习, 并被广泛研究 (参见文献 [34, 54]). 通过引入多智能体系统、图神经网络等新技术, 给建模多个细胞间的交互机制提供新思路. 但这些机器学习方法均基于跟踪数据, 而时间序列单细胞数据会丢失不同时间点间细胞的关联性. 未来, 如何更好地结合细胞间通信、整合不同时间点的异质数据, 将是该领域进一步发展的关键.

近年来, 时空数据逐渐涌现, 为生物系统建模提供了更多的维数和信息. 当前, 已经有一些时空模型开始尝试将时间序列数据与空间数据有效结合 [26, 35, 44]. 例如, Huizing 等 [26] 基于融合 Gromov-Wasserstein 距离将空间信息嵌入损失函数, 使得模型不仅能够在基因表达空间上匹配细胞分布, 还能在细胞的空间位置上进行有效的匹配. Peng 等 [44] 基于非平衡最优传输理论, 提出了 stVCR 方法. stVCR 将基因表达与空间信息同时建模, 为时空尺度上的生物系统建模提供了一种新的工具. 在未来的研究中, 如何将时间、空间与数据结合, 从动态和随机的视角来分析和模拟细胞在时空维数上的行为, 将成为研究的关键方向.

总体而言, 动态建模与深度学习的交叉领域, 不仅可以通过引入新的计算工具, 深化对复杂生物系统的理解, 还可能在推动数学理论, 尤其是概率论、随机过程及最优传输理论的发展方面发挥重要作用. 未来, 随着跨学科合作的加强, 生物学与数学的深度融合必将催生出更多创新性成果和应用, 为精准医疗和生物学研究开辟新的方向.

参考文献

1 Armingol E, Officer A, Harismendy O, et al. Deciphering cell-cell interactions and communication from gene expression. Nat Rev Genet, 2021, 22: 71–88
2 Benamou J-D, Brenier Y. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. Numer Math, 2000, 84: 375–393
3 Benamou J-D, Carlier G, Laborde M. An augmented Lagrangian approach to Wasserstein gradient flows and applications. ESAIM Proc, 2016, 54: 1–17
4 Bunne C, Hsieh Y-P, Cuturi M, et al. Recovering stochastic dynamics via Gaussian Schrödinger bridges. arXiv: 2202.05722, 2022
5 Bunne C, Papaxanthos L, Krause A, et al. Proximal optimal transport modeling of population dynamics. In: Proceedings of the 25th International Conference on Artificial Intelligence and Statistics. San Diego: JMLR, 2022, 6511–652
6 Bressloff P C. Stochastic Processes in Cell Biology, 2nd ed. Heidelberg: Springer, 2021
7 Browaeys R, Saelens W, Saeys Y. NicheNet: Modeling intercellular communication by linking ligands to target genes. Nat Methods, 2020, 17: 159–162
8 Bryson A E, Ho Y. Applied Optimal Control: Optimization, Estimation and Control. Boca Raton: CRC Press, 1975
9 Carrillo J A, Craig K, Wang L, et al. Primal dual methods for Wasserstein gradient flows. Found Comput Math, 2021, 22: 1–55
10 Chen F, Li C. Inferring structural and dynamical properties of gene networks from data with deep learning. NAR Genomics Bioinform, 2022, 4: lqac068
11 Chen R T Q, Rubanova Y, Bettencourt J, et al. Neural ordinary differential equations. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2018, 6572–6583
12 Chen X, Jia C. Limit theorems for generalized density-dependent Markov chains and bursty stochastic gene regulatory networks. J Math Biol, 2020, 80: 959–994
13 Chen Z W, An S K, Bai X Q, et al. DensityPath: An algorithm to visualize and reconstruct cell state-transition path on density landscape for single-cell RNA sequencing data. Bioinformatics, 2019, 35: 2593–2601
14 Chizat L, Peyré G, Schmitzer B, et al. An interpolating distance between optimal transport and Fisher-Rao metrics. Found Comput Math, 2018, 18: 1–44
15 Chow S-N, Li W C, Zhou H M. Entropy dissipation of Fokker-Planck equations on graphs. Discrete Continuous Dyn Syst Ser A, 2018, 38: 4929–4950
16 Cuturi M. Sinkhorn Distances: Lightspeed computation of optimal transport. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2013, 2292–2300
17 E W N , Li T J, Vanden-Eijnden E. Applied Stochastic Analysis. Graduate Studies in Mathematics, vol. 199. New York: Amer Math Soc, 2021
18 Efremova M, Vento-Tormo M, Teichmann S A, et al. CellPhoneDB: Inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. Nat Protoc, 2020, 15: 1484–1506
19 Fang X N, Kruse K, Lu T, et al. Nonequilibrium physics in biology. Rev Modern Phys, 2019, 91: 045004
20 Gao J X, Wang Z Y, Zhang M Q, et al. A data-driven method to learn a jump diffusion process from aggregate biological gene expression data. J Theoret Biol, 2022, 532: 110923
21 Ge H, Qian H. Analytical mechanics in stochastic dynamics: Most probable path, large-deviation rate function and Hamilton-Jacobi equation. Internat J Modern Phys B, 2012, 26: 1230012
22 Goldstein H, Poole C, Safko J. Classical Mechanics, 3rd ed. New York: Pearson Education, 2011
23 Grathwohl W, Chen R T Q, Bettencourt J, et al. FFJORD: Free-form continuous dynamics for scalable reversible generative models. In: Proceedings of the 7th International Conference on Learning Representations. New Orleans: ICLR, 2019, https://openreview.net/forum?id=rJxgknCcK7
24 Höfer F, Soner H M. Potential mean-field games and gradient flows. arXiv:2408.00733, 2024
25 Huguet G, Magruder DS, Tong A, et al. Manifold interpolating optimal-transport flows for trajectory inference. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2022, 29705–29718
26 Huizing G J, Peyre G, Cantini L. Learning cell fate landscapes from spatial transcriptomics using Fused Gromov-Wasserstein. bioRxiv, 2024, doi: https://doi.org/10.1101/2024.07.26.605241

27 Jia C, Qian M P, Kang Y, et al. Modeling stochastic phenotype switching and bet-hedging in bacteria: Stochastic nonlinear dynamics and critical state identification. Quant Biol, 2014, 2: 110–125
28 Jiang Q, Wan L. A physics-informed neural SDE network for learning cellular dynamics from time-series scRNA-seq data. Bioinformatics, 2024, 40: 120–127
29 Jiang Q, Zhang S, Wan L. Dynamic inference of cell developmental complex energy landscape from time series single-cell transcriptomic data. PLoS Comput Biol, 2022, 18: e1009821
30 Jin S Q, Guerrero-Juarez C F, Zhang L H, et al. Inference and analysis of cell-cell communication using CellChat. Nat Commun, 2021, 12: 1088
31 Jordan R, Kinderlehrer D, Otto F. The variational formulation of the Fokker-Planck equation. SIAM J Math Anal, 1998, 29: 1–17
32 Kang X, Li C H. A dimension reduction approach for energy landscape: Identifying intermediate states in metabolism-EMT network. Adv Sci, 2021, 8: 2003133
33 Kantorovich L V. On the translocation of masses. Manag Sci, 1958, 5: 1–4
34 Kipf T, Fetaya E, Wang K-C, et al. Neural relational inference for interacting systems. In: Proceedings of the 35th International Conference on Machine Learning. San Diego: JMLR, 2018, 2688–2697
35 Klein D, Palla G, Lange M, et al. Mapping cells through time and space with moscot. bioRxiv, 2023, doi: https://doi.org/10.1101/2023.05.11.540374
36 Li W C. A study of stochastic differential equations and Fokker-Planck equations with applications. PhD Thesis. Atlanta: Georgia Institute of Technology, 2016
37 Li X C, Wong T-K L, Chen R T Q, et al. Scalable gradients for stochastic differential equations. In: Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics. Boston: Addison-Wesley Publ, 2020, 3870–3882
38 Lv C, Li X, Li F T, et al. Constructing the energy landscape for genetic switching system driven by intrinsic noise. PLoS ONE, 2014, 9: e88167
39 Monge G. Mémoire sur la théorie des déblais et des remblais. In: Mémoires de l'Académie Royale des Sciences. Paris: Imprimerie Royale, 1781, 666–704
40 Muzellec B, Cuturi M. Subspace detours: Building transport plans that are optimal on subspace projections. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2019, 6917–6928
41 Onken D, Fung S W, Li X J, et al. OT-Flow: Fast and accurate continuous normalizing flows via optimal transport. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2021, 35: 9223–9232
42 Otto F. The geometry of dissipative evolution equations: The porous medium equation. Comm Partial Differential Equations, 2001, 26: 101–174
43 Park C, Mani S, Beltran-Velez N, et al. A Bayesian framework for inferring dynamic intercellular interactions from time-series single-cell data. Genome Res, 2024, 34: 1384–1396
44 Peng Q W, Zhou P J, Li T J. stVCR: Reconstructing spatio-temporal dynamics of cell development using optimal transport. bioRxiv, 2024, doi: 10.1101/2024.06.02.596937-06
45 Peyré G. Entropic approximation of Wasserstein gradient flows. SIAM J Imag Sci, 2015, 8: 2323–2351
46 Peyré G, Cuturi M. Computational optimal transport: With applications to data science. Found Trends Mach Learn, 2019, 11: 355–607
47 Rubanova Y, Chen R T Q, Duvenaud D. Latent ordinary differential equations for irregularly-sampled time series. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2019, 5320–5330
48 Ruthotto L, Osher S J, Li W C, et al. A machine learning framework for solving high-dimensional mean field game and mean field control problems. Proc Natl Acad Sci USA, 2020, 117: 9183–9193
49 Schiebinger G, Shu J, Tabaka M, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. Cell, 2019, 176: 928–943
50 Sha Y T, Qiu Y C, Zhou P J, et al. Reconstructing growth and dynamic trajectories from single-cell transcriptomics data. Nat Mach Intell, 2024, 6: 25–39
51 Sinkhorn R. A relationship between arbitrary positive matrices and doubly stochastic matrices. Ann of Math Stud, 1964, 35: 876–879
52 Smart M, Zilman A. Emergent properties of collective gene-expression patterns in multicellular systems. Cell Rep Phys Sci, 2023, 4: 101247
53 Street K, Risso D, Fletcher R B, et al. Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics, 2018, 19: 477
54 Sun F Y, Kauvar I, Zhang R H, et al. Interaction modeling with multiplex attention. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2022, 20038–20050
55 Tanay A, Regev A. Scaling single-cell genomics from phenomenology to mechanism. Nature, 2017, 541: 331–338

56 Tong A, Huang J, Wolf G, et al. TrajectoryNet: A dynamic optimal transport network for modeling cellular dynamics. Proc Mach Learn Res, 2020, 119: 9526–9536

57 Trapnell C, Cacchiarelli D, Grimsby J, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol, 2014, 32: 381–386

58 Vargas F, Thodoroff P, Lamacraft A, et al. Solving Schrödinger bridges via maximum likelihood. Entropy, 2021, 23: 1134

59 Villani C. Optimal Transport: Old and New. Berlin: Springer, 2009

60 Waddington C H. The Strategy of the Genes: A Discussion of Some Aspects of Theoretical Biology. London: G. Allen and Unwin, 1957

61 Wang J. Landscape and flux theory of non-equilibrium dynamical systems with application to biology. Adv Phys, 2015, 64: 1–137

62 Wang J, Xu L, Wang E. Potential landscape and flux framework of nonequilibrium networks: Robustness, dissipation, and coherence of biochemical oscillations. Proc Natl Acad Sci USA, 2008, 105: 12271–12276

63 Wolf F A, Hamey F K, Plass M, et al. PAGA: Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol, 2019, 20: 59

64 Yang K D, Uhler C. Scalable unbalanced optimal transport using generative adversarial networks. In: Proceedings of the 7th International Conference on Learning Representations. New Orleans: ICLR, 2019, https:// openreview.net/forum?id=HyexAiA5Fm

65 Yang L, Daskalakis C, Karniadakis G E. Generative ensemble regression: Learning particle dynamics from observations of ensembles with physics-informed deep generative models. SIAM J Sci Comput, 2022, 44, B80–B99

66 Yeo G H T, Saksena S D, Gifford D K. Generative modeling of single-cell time series with prescient enables prediction of cell trajectories with interventions. Nat Commun, 2021, 12: 3222

67 Zhang J Q, Larschan E, Bigness J, et al. scNODE: Generative model for temporal single cell transcriptomic data prediction. Bioinformatics, 2024, 40: 146–154

68 Zhang K, Zhu J H, Kong D H, et al. Modeling single cell trajectory using forward-backward stochastic differential equations. PLoS Comput Biol, 2024; 20: e1012015

69 Zhang Z Y, Li T J, Zhou P J. Learning stochastic dynamics from snapshots through regularized unbalanced optimal transport. arXiv:2410.00844, 2024

70 Zhong Y D, Biswadip D, Amit C. Extending lagrangian and hamiltonian neural networks with differentiable contact models. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2021, 21910–21922

Dynamic modeling, optimization, and deep learning for high-dimensional complex biological data

Qi Jiang & Lin Wan

Abstract The continuous emergence of high-dimensional biological big data presents new challenges for mathematical modeling and computation in the fields of probability and statistics, data science, and artificial intelligence. In this paper, we review the research progress in dynamic modeling and computation of snapshot-type high-dimensional data, with time-series single-cell sequencing data as a representative example. We focus on mathematical models and optimization methods based on optimal transport theory, emphasizing the key role of integrating mathematical models with deep learning for efficient computation in high-dimensional big data. Additionally, we discuss the challenges and opportunities in modeling and computation of complex nonlinear systems, such as inferring cell-cell interactions.

Keywords **snapshot-type high-dimensional data, dynamic modeling, optimal transport, optimal control, deep learning**

MSC(2020) **60J70, 49Q22, 37N25, 92C37**

doi: 10.1360/SSM-2024-0308