SIMULATION-FREE STRUCTURE LEARNING FOR STOCHASTIC DYNAMICS

Noah El Rimawi-Fine[1,2*] **Adam Stecklov**[1,2*] **Lucas Nelson**[1,2*] **Alexander Tong**[2,3]
Mathieu Blanchette[1,2] **Stephen Y. Zhang**[4] **Lazar Atanackovic**[5,6]
[1]McGill University [2]Mila – Quebec AI Institute [3]Université de Montréal
[4]University of Melbourne [5]University of Toronto [6]Vector Institute

Introduction. Many physical systems, notably in cell and molecular biology, operate in an out-of-equilibrium regime and are subject to intrinsic stochasticity. Unravelling the structure of the underlying system from noisy, partial measurements is a central scientific problem that is key to gaining mechanistic understanding and predictive insights into system behaviour. We consider a generic class of processes, such as gene expression dynamics in cells, which can be modelled using a (possibly high-dimensional) continuous stochastic dynamical system of the form

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{v}(\boldsymbol{X}_t)\,\mathrm{d}t + \boldsymbol{\sigma}\,\mathrm{d}\boldsymbol{B}_t, \quad \boldsymbol{X} \in \mathbb{R}^d, \tag{1}$$

where $\boldsymbol{v}$ models an (a priori unknown) *vector field* describing deterministic aspects of the system, and $\boldsymbol{\sigma}$ determines the noise intensity. Given some set of observations on $\boldsymbol{X}_t$ drawn from equation 1, two related tasks arise. The first task, which we refer to as **dynamical inference**, deals with constructing from measurements an estimate of the underlying vector field $\boldsymbol{v}$. From a good estimate of $\boldsymbol{v}$, properties of the dynamics such as bifurcations or attractors can be deduced. On the other hand, **structure learning** aims to reconstruct the *causal* relationships between each of the d variables in $\boldsymbol{X}$. A system with only dyadic causal relationships can be described thus by a directed graph with d vertices, where an edge $i \to j$ implies that variable i "causes" j in some appropriate sense[1]. Naturally, some knowledge of $\boldsymbol{v}$ is advantageous for this task, also referred in various literatures as a *network inference*, *structure learning*, and *system identification*.

We consider the setting where our observations consist of *snapshots* of particles from equation 1 at a series of *time-points* $t_1, \ldots, t_T$. Concretely, we observe the following empirical distributions for $1 \leq i \leq T$, $\hat{\boldsymbol{p}}_{t_i} = N_i^{-1} \sum_{j=1}^{N_i} \delta_{\boldsymbol{X}_j^{(i)}}$, where $\{\boldsymbol{X}_j^{(i)}\}_j$ are independent samples from $\mathbb{P}(\boldsymbol{X}_{t_i} = \cdot)$, and each timepoint is sampled from an independent realization of the process equation 1. We aim to jointly learn a *functional* reconstruction of the vector field $\boldsymbol{v}$ as well as a representation of the causal dependencies between the d variables at play in the form of a $d \times d$ graph G.

We build upon a number of recent works [2, 1, 3] and introduce a principled framework for the joint estimation of $\boldsymbol{v}$ as well as the graph G. We use simulation-free flow and score matching (SF2M) [2] conditioned on entropy-regularized optimal transport (EOT) couplings to regress $\hat{\boldsymbol{v}}$. We parameterize $\hat{\boldsymbol{v}}$ with a neural graphical model (NGM) [1] which incorporates the causal graph G into its underlying network structure. Our approach (NGM-SF2M)[2] is simulation-free during training, which enables scaling to high-dimensional systems compared to the counterpart approaches.

Method. Given a pair of snapshot observations $\hat{\boldsymbol{p}}_{t_i}, \hat{\boldsymbol{p}}_{t_{i+1}}$, solving the EOT problem

$$\min_{\pi \text{ a coupling of } \hat{\boldsymbol{p}}_{t_i}, \hat{\boldsymbol{p}}_{t_{i+1}}} \langle C, \pi \rangle + \varepsilon \langle \pi, \log \pi \rangle \tag{2}$$

with cost $C(x, y) = \frac{1}{2}\|x - y\|_2^2$ and $\varepsilon = \sigma^2(t_{i+1} - t_i)$ yields the *static* Schrödinger Bridge (SB) between the two snapshots. When $\boldsymbol{v} = -\nabla V$, this approach is known to be consistent in recovering the vector field [4]. We thus use the SB (equation 2) as the building block for our estimator. equation 2 is associated with a corresponding vector field $\boldsymbol{u}$ that generates the *dynamic* SB between t_i and t_{i+1}.

SF2M uses the elegant connection of EOT with the dynamic SB along with characterizations of the Brownian bridge to provide a simulation-free approach to learning a neural estimator of $\boldsymbol{u}$. From equation 1 we seek to recover the *autonomous* field $\boldsymbol{v}$, while the objective in [2] targets the SB probability *flow* $\boldsymbol{u}$. The quantities of the ground truth process equation 1 are related via

*Joint first authorship. Correspondence to: l.atanackovic@mail.utoronto.ca

[1]We remark that numerous, non-equivalent, definitions of causality (Granger, transfer entropy, etc.) exist. Here, we consider as our criterion *local independence* [1] which is well-defined for continuous-time systems.

[2]We note that [2] introduces an approach for joint structure learning and dynamical inference, but is limited to unconditional dynamics. Our method differs in several ways, most notably in that we incorporate interventional data and learn conditional dynamics, improving the structure learning and dynamical inference performance.

$\boldsymbol{v}_{\text{flow}}(x) = \boldsymbol{v}(x) - \frac{\sigma^2}{2}\nabla \log p_t(x)$, where $u_t, \nabla \log p_t$ are the probability flow field and the score of the Brownian bridge pinned at $(t_i, x_i), (t_{i+1}, x_{i+1})$ respectively. We therefore make the parametrization choice $\boldsymbol{u}_t(x) = \hat{\boldsymbol{v}}^\theta(x) - \frac{\sigma^2}{2}\boldsymbol{s}_t^\phi(x)$.

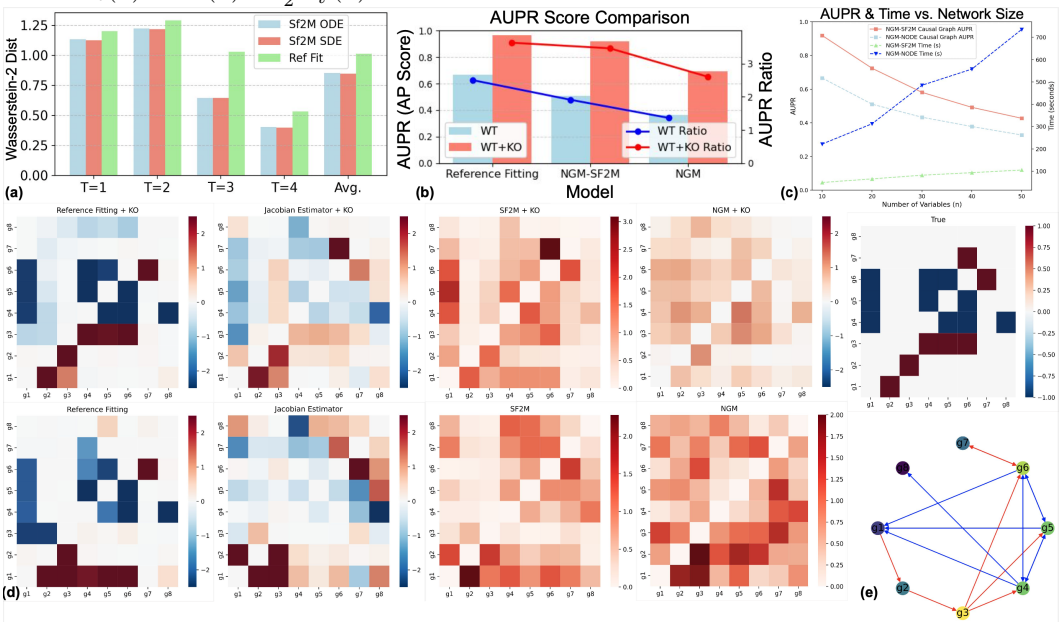

Figure 1: (a) Dynamical inference accuracy (b) AUPR Score Comparison (c) NGM-SF2M vs. NGM-NODE scaling efficiency (d) Jacobian and Causal Network Heatmaps (e) Network Construction

Results. We consider a synthetic dataset following the paradigm in [3]. We use BoolODE [5] to simulate gene expression dynamics for a system with 8 genes ($d = 8$) which exhibits trifurcating trajectories. We do this for the observational response (wild-type, no interventions) and response under 7 different interventions (knockouts) where we remove the respective gene from the system. After simulation, we construct 5 marginals with 1,000 cells each in the observational and interventional settings. We compare with two methods which consider joint inference of underlying network structure and system dynamics – Reference Fitting (RF) [3] and nonlinear models such as the NGM which requires simulation during training (trained via NeuralODEs) [1].

We show our results in Figure 1. We observe that using interventions (knockouts) significantly improves the network inference accuracy for all models, with observable improvements in AUPR and gene-gene interaction strength Figure 1(b)(d). Although RF outperforms NGM-SF2M, this is likely due to the linearization process which captures the dominant trends in the transient dynamics, which were sufficient to explain the key network interactions. To evaluate the performance of NGM-SF2M for dynamical inference, we use a leave-one-out time point test set, where models were trained excluding time point t and then used to simulate trajectories from $t-1$ to t. Performance was measured using the Wasserstein-2 distance between simulated and ground truth distributions . As shown in Figure 1(a), NGM-SF2M outperforms RF across all leave-one-out timepoints, with the SDE variant achieving the lowest Wasserstein-2 distance. Lastly, we evaluate the performance of NGM-SF2M (simulation-free) relative to NGM (requires simulation) as d increases. For this, we consider a toy linear system where we can control d. From Figure 1(c), we observe that NGM-SF2M can maintain competitive AUPR relative to NGM-NODE at the fraction of the computational cost.

Conclusion. We introduce a principled approach for jointly recovering the underlying network structure and dynamic response of a physical system. We show that our simulation-free method, NGM-SF2M, not only exhibits improved scaling relative to NGM on progressively larger linear systems, but also consistently retrieves a competitive recovery of the underlying network structure. Moreover, we show that incorporating interventional data yields improved performance for inferring network interactions. Our results indicate that while RF recovers marginally more accurate network structure, NGM-SF2M yields improved performance on the joint task – dynamical inference and structure learning. In future work, we aim to extend our framework to higher-dimensional systems, real-world settings, and integrate multi-modal data such as chromatin accessibility.

Meaningfulness. The stochastic dynamical systems we consider here are used ubiquitously in the biophysics and modelling literature. In a biological context, our vector field $\boldsymbol{v}$ can be related to the metaphorical "Waddington's landscape".

Acknowledgments. The authors acknowledge funding from UNIQUE, CIFAR, NSERC, Intel, and Samsung. The research was enabled in part by computational resources provided by the Digital Research Alliance of Canada (https://alliancecan.ca), Mila (https://mila.quebec), the Province of Ontario, companies sponsoring the Vector Institute (http://vectorinstitute.ai/partners/), and NVIDIA

REFERENCES

[1] Alexis Bellot, Kim Branson, and Mihaela van der Schaar. "Neural Graphical Modelling in Continuous-Time: Consistency Guarantees and Algorithms". In: *International Conference on Learning Representations* (2022).

[2] Alexander Tong et al. "Simulation-free Schrödinger Bridges via Score and Flow Matching". In: *International Conference on Artificial Intelligence and Statistics* (2024).

[3] Stephen Y. Zhang. "Joint Trajectory and Network Inference via Reference Fitting". In: *Proceedings of the 19th Machine Learning in Computational Biology meeting* (2024).

[4] Hugo Lavenant, Stephen Zhang, Young-Heon Kim, and Geoffrey Schiebinger. "Towards a mathematical theory of trajectory inference". In: *The Annals of Applied Probability* (2024).

[5] Aditya Pratapa, Amogh P Jalihal, Jeffrey N Law, Aditya Bharadwaj, and TM Murali. "Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data". In: *Nature Methods* 17.2 (2020), pp. 147–154. DOI: 10.1038/s41592-019-0690-6.