GOTFlow: Learning Directed Population Transitions from Cross-Sectional Biomedical Data with Optimal Transport

George Wright[1], Ethar Alzaid[1], Joanne Muter[2], Jan Brosens[2], Fayyaz Minhas[1,2]
[1] Predictive Systems in Biomedicine (PRISM) Lab, Department of Computer Science
[2] Warwick Medical School
University of Warwick, Coventry, United Kingdom (Corresponding authors: george.wright.1@warwick.ac.uk, ethar.alzaid@warwick.ac.uk, fayyaz.minhas@warwick.ac.uk)

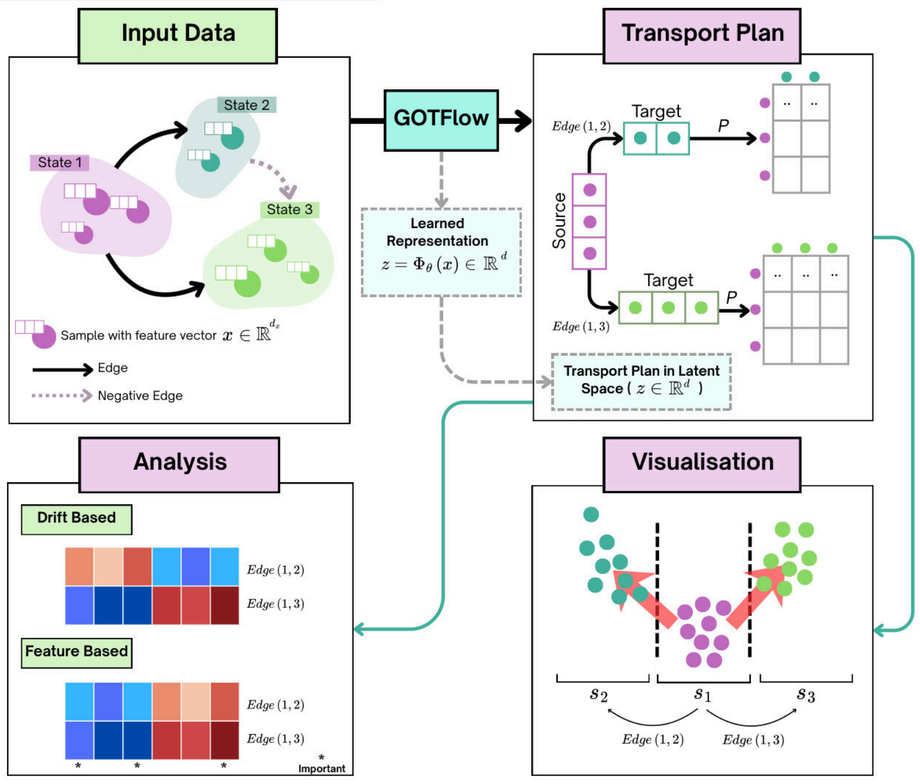
Predictive Systems in Biomedicine Lab PRISM

UNIVERSITY OF WARWICK

Motivation

- *Biology unfolds as trajectories. Patient cohorts are often only snapshots.*
- *Rather than inferring a trajectory from data alone, can we test a hypothesised graph of biological transitions and determine which transitions are supported, how they occur, and what drives them?*

Method Overview



What are the inputs?

- A dataset $(x_i, s_i)_{i=1}^N$, with features $x_i \in \mathbb{R}^{d_x}$ and state labels $s_i \in \{1, .., K\}$.
- A directed transition graph $G = \{(S_l, T_l)\}_{l=1}^L$, each edge is a hypothesised transition from state S to T.

What is learned?

A latent representation and unbalanced Optimal Transport (UOT) couplings that make admissible transitions minimum-cost, learned contrastively against implausible transitions.

Transition energy for edge (S, T):

$$E_{ST}(\theta) = \min_{P \geq 0} \mathcal{U}(\tilde{\mu}_S, \tilde{\mu}_T; P) - \frac{1}{2} \sum_{X \in \{S,T\}} \min_{P \geq 0} \mathcal{U}(\tilde{\mu}_X, \tilde{\mu}_X; P)$$

Training objective (with contrastive loss $\mathcal{L}_{NCE}$):

$$\min_\theta \mathcal{L}(\theta) = \sum_{(S,T) \in G} w_{ST} \left[E_{ST}(\theta) + \gamma \mathcal{L}_{\text{NCE}}^{(S,T)}(\theta) \right] + \Omega(\theta)$$

What is obtained?

- **A transport plan *P*** coupling each source sample to its targets in the next state.
- **A latent mapping** Φ in which admissible transitions are minimum-cost.
- **Drift vectors** summarising each sample's transition, which back-project to feature space to reveal transition drivers.
- **Transition energies** E_{ST}, scoring the cost of each hypothesised transition.

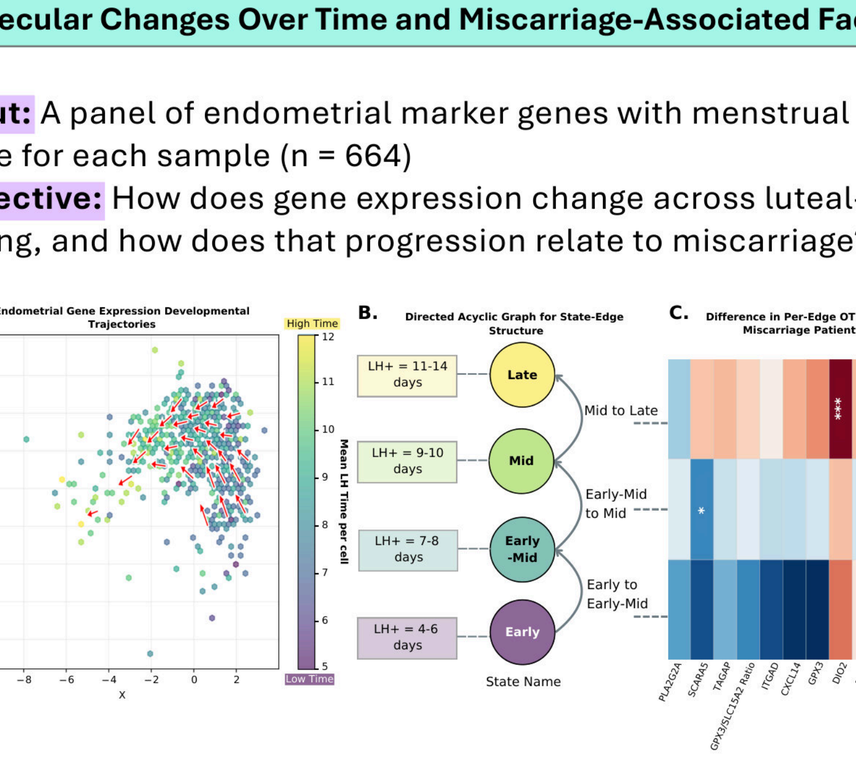
Key Take Aways

- GOTFlow can reconstruct directed population transitions from cross-sectional data, guided by a hypothesised transition graph rather than an assumed linear trajectory.
- Quantifies feature-level changes and transition energies, revealing both the transition drivers and support for each proposed transition.

Validation and Biomedical Applications

Recovery of Known Branching and Merging Transitions

Input: 100 samples in 64-D with known ground-truth branching and merging transitions.



Findings: GOTFlow recovers the correct directional structure (mean cosine similarity 0.720 ± 0.152 between inferred and true drift; chance ≈ 0).

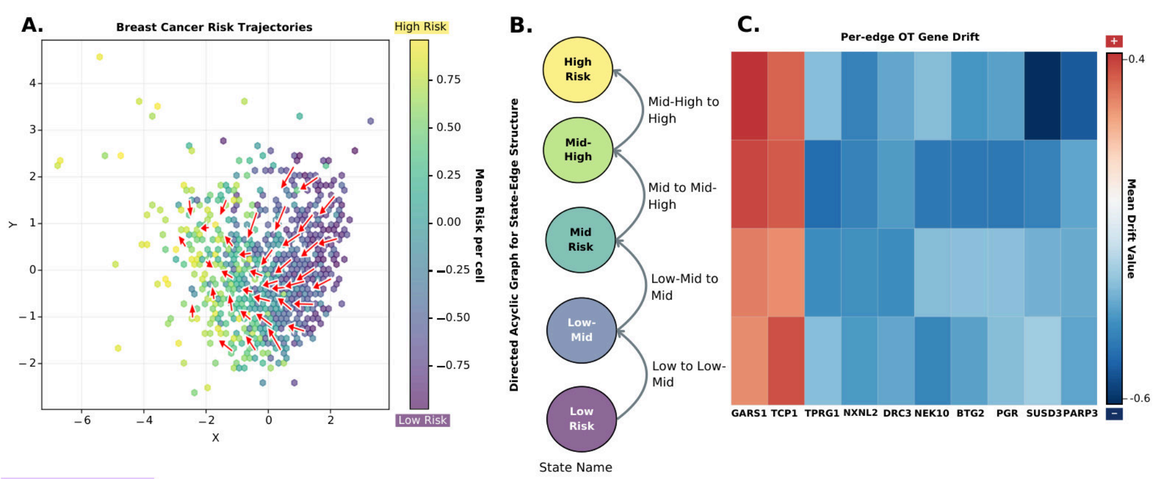
Molecular Changes Over Time and Miscarriage-Associated Factors

Input: A panel of endometrial marker genes with menstrual timing state for each sample (n = 664)
Objective: How does gene expression change across luteal-phase timing, and how does that progression relate to miscarriage?

Findings: Miscarriage samples show reduced inferred forward drift across luteal timing states, consistent with impaired decidualisation.

Inferring Molecular Changes Across Breast Cancer Risk States

Input: Patient-level expression profiles of prognostic genes (n=1061), with each patient assigned to a known survival-risk state (low → high).
Objective: How does gene expression change across increasing survival-risk states, and which genes drive that shift?



Findings: Coherent low-to-high-risk flow, with risk-driving genes rising (TCP1, GARS1) and protective ones falling (SUSD3, NXNL2).

References

[1] Muter, J. et al . Stalling of the endometrial decidual reaction determines the recurrence risk of miscarriage. Science Advances, 2025.
[2] J Liu et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell, 2018.

Code

Full Paper