Causal Discovery: Revealing Hidden Patterns in Biology with Machine Learning

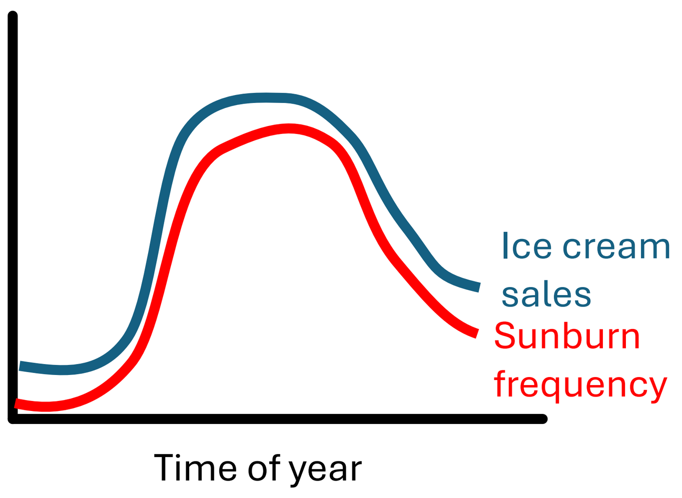
Holly Chambers[1], Hervé Isambert[2], Vahid Shahrezaei[1], Barbara Bravi[1]

[1]Department of Mathematics, Imperial College London
[2]CNRS UMR168, Institut Curie, Université PSL, Sorbonne Université

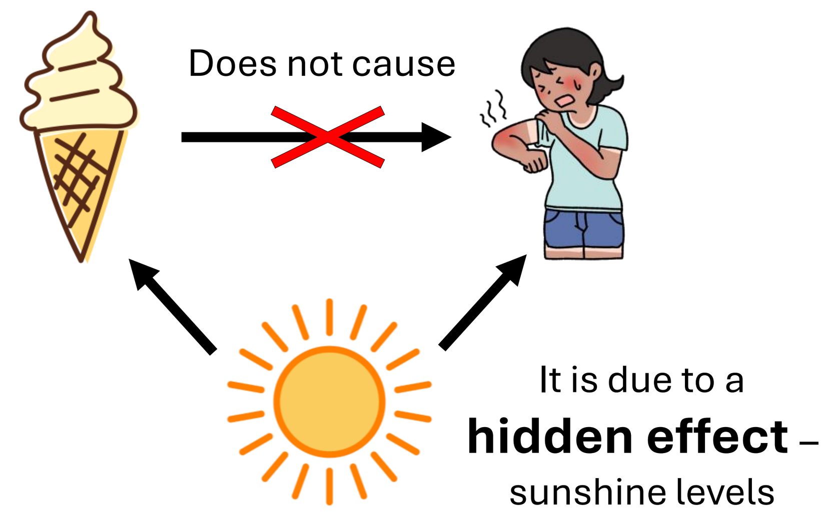
What is Causal Discovery?

Causal discovery uses data to determine which factors influence others. It allows us to separate **cause** from **coincidence**.

For example, when ice cream sales increase, so do rates of sunburn



This is not because ice cream causes sunburn



We apply these methods to biological data where identifying hidden effects is important for understanding and developing treatments for disease. If we were trying to reduce rates of sunburn, it is important to know that stopping ice cream sales won't help.

How to Build a Causal Network

Step 1 – Collect data

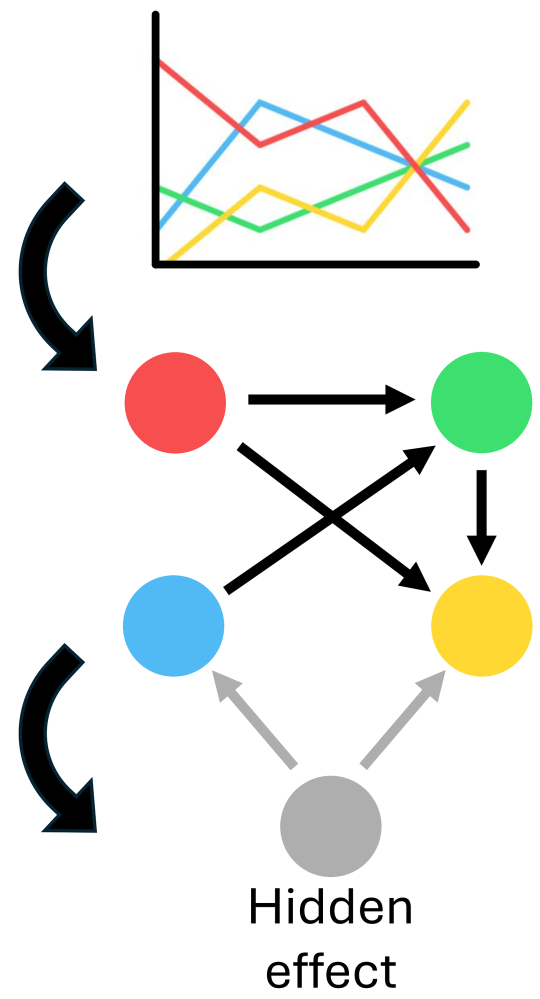
Gather biological data such as gene expression, protein levels, or patient symptoms.

Step 2 – Find patterns

Identify variables that tend to change together.

Step 3 – Apply causal inference

Use mathematical techniques to test if linked variables cause one another, or if a hidden effect is involved.

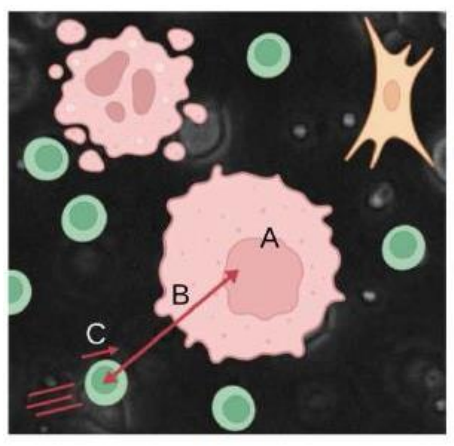


A link → means that causes

We test the ability of tMIIC[1] (temporal Multivariate Information-based Inductive Causation) to reconstruct known networks from time series data.

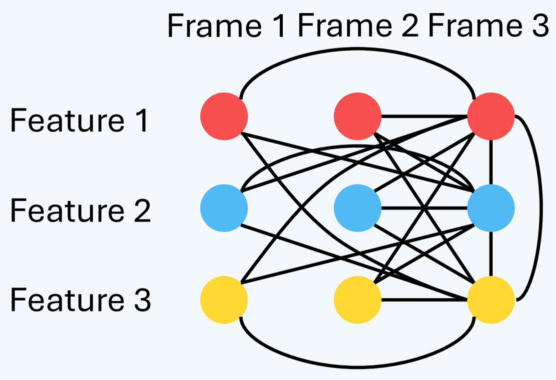
tMIIC: Causal Discovery for Time Series Data

Imagine an experiment tracking some cells, taking a video to record how features such as movement, shape and division change over time. Each feature is a **node**, and each frame of the video is a **time step**.



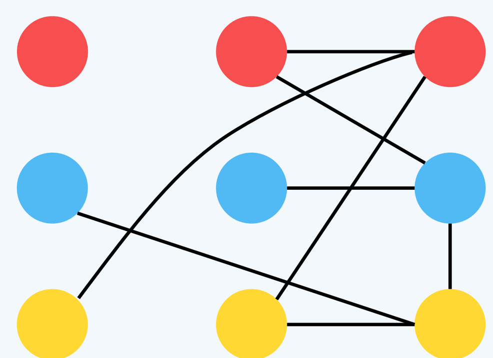
A: Cell area
B: Cell proximity
C: Cell speed

Step 1: Fully connected network



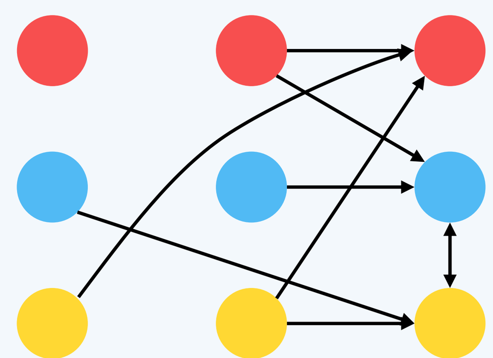
Any feature could potentially influence any other.

Step 2: Remove unnecessary links



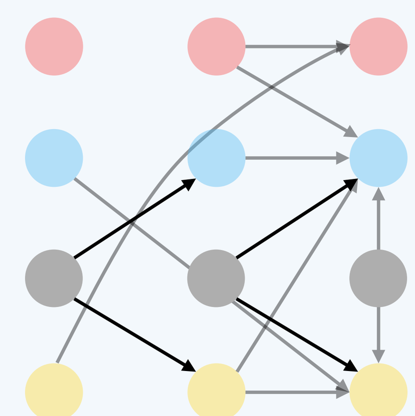
If removing does not change the predictability of .

Step 3: Add arrows to show direction of causal influence



All causal links must point forward in time.

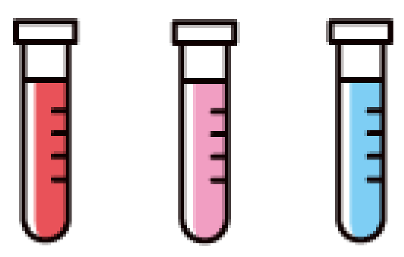
Step 4: Infer hidden effects



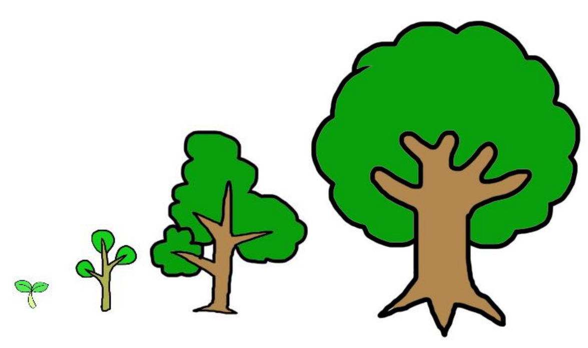
Two variables are linked, but no with direct cause: a bi-directed edge is added.

Improving Causal Discovery

Our work shows that temporal causal **discovery works best when data is collected at appropriate intervals**. Infrequent measurements mean important details are missed. Too frequent, and noise may obscure meaningful patterns.



Chemical reactions: very **fast**, take measurements over seconds



Tree growth: very **slow**, take measurements over years

The optimal sampling rate for the data depends on the **timescale**. We use statistical methods to calculate this and choose the best sampling rate for tMIIC.

Conclusions

Identifying causal relationships is important for **understanding complex biological systems** and can provide a starting point for mathematical models.

Our work has shown that **causal discovery methods are effective at this task, and their performance can be optimized**, for example by inputting data taken at intervals that best capture the dynamics of the system.

References

1 Simon, Franck, et al. "CausalXtract, a flexible pipeline to extract causal effects from live-cell time-lapse imaging data." *eLife* 13 (2025): RP95485.

IMPERIAL

h.chambers22@imperial.ac.uk