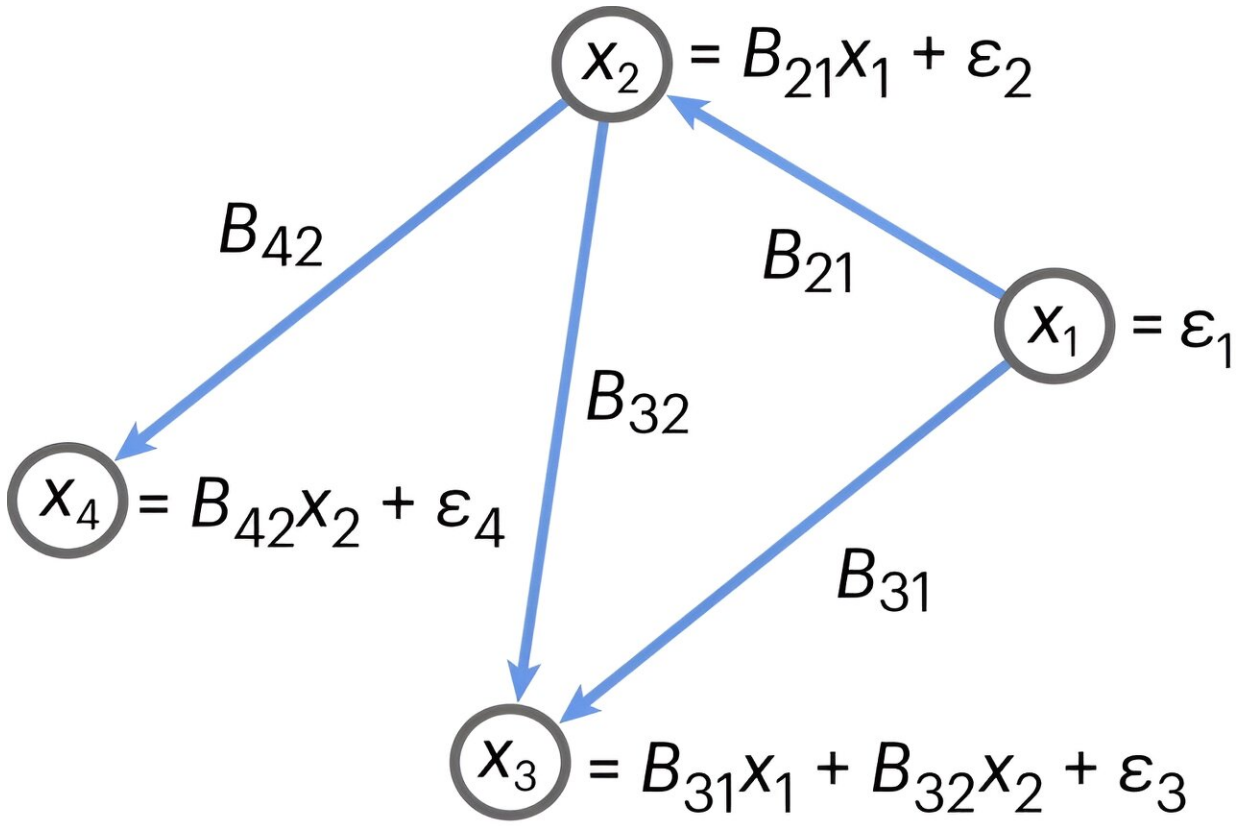
A more effective experimental design for engineering a cell into a new state

October 2 2023, by Adam Zewe



Example causal model. An example of a linear SCM with Gaussian noise on a four-node DAG, in which nodes and edges are labeled with variables and coefficients, respectively. Credit: *Nature Machine Intelligence* (2023). DOI: 10.1038/s42256-023-00719-0

A strategy for cellular reprogramming involves using targeted genetic

interventions to engineer a cell into a new state. The technique holds great promise in immunotherapy, for instance, where researchers could reprogram a patient's T-cells so they are more potent cancer killers. Someday, the approach could also help identify life-saving cancer treatments or regenerative therapies that repair disease-ravaged organs.

But the human body has about 20,000 genes, and a genetic perturbation could be on a combination of genes or on any of the over 1,000 transcription factors that regulate the genes. Because the search space is vast and genetic experiments are costly, scientists often struggle to find the ideal perturbation for their particular application.

Researchers from MIT and Harvard University developed a new, computational approach that can efficiently identify optimal genetic perturbations based on a much smaller number of experiments than traditional methods.

Their algorithmic technique leverages the cause-and-effect relationship between factors in a complex system, such as genome regulation, to prioritize the best intervention in each round of sequential experiments.

The researchers conducted a rigorous theoretical analysis to determine that their technique did, indeed, identify optimal interventions. With that theoretical framework in place, they applied the algorithms to real biological data designed to mimic a cellular reprogramming experiment. Their algorithms were the most efficient and effective.

"Too often, large-scale experiments are designed empirically. A careful causal framework for sequential experimentation may allow identifying optimal interventions with fewer trials, thereby reducing experimental costs," says co-senior author Caroline Uhler, a professor in the Department of Electrical Engineering and Computer Science (EECS) who is also co-director of the Eric and Wendy Schmidt Center at the

Broad Institute of MIT and Harvard, and a researcher at MIT's Laboratory for Information and Decision Systems (LIDS) and Institute for Data, Systems and Society (IDSS).

Joining Uhler on the paper, which appears today in *Nature Machine Intelligence*, are lead author Jiaqi Zhang, a graduate student and Eric and Wendy Schmidt Center Fellow; co-senior author Themistoklis P. Sapsis, professor of mechanical and ocean engineering at MIT and a member of IDSS; and others at Harvard and MIT.

Active learning

When scientists try to design an effective intervention for a complex system, like in cellular reprogramming, they often perform experiments sequentially. Such settings are ideally suited for the use of a machine-learning approach called active learning. Data samples are collected and used to learn a model of the system that incorporates the knowledge gathered so far. From this model, an acquisition function is designed—an equation that evaluates all potential interventions and picks the best one to test in the next trial.

This process is repeated until an optimal intervention is identified (or resources to fund subsequent experiments run out).

"While there are several generic acquisition functions to sequentially design experiments, these are not effective for problems of such complexity, leading to very slow convergence," Sapsis explains.

Acquisition functions typically consider correlation between factors, such as which genes are co-expressed. But focusing only on correlation ignores the regulatory relationships or causal structure of the system. For instance, a genetic intervention can only affect the expression of downstream genes, but a correlation-based approach would not be able

to distinguish between genes that are upstream or downstream.

"You can learn some of this causal knowledge from the data and use that to design an intervention more efficiently," Zhang explains.

The MIT and Harvard researchers leveraged this underlying causal structure for their technique. First, they carefully constructed an algorithm so it can only learn models of the system that account for causal relationships.

Then the researchers designed the acquisition function so it automatically evaluates interventions using information on these causal relationships. They crafted this function so it prioritizes the most informative interventions, meaning those most likely to lead to the optimal intervention in subsequent experiments.

"By considering causal models instead of correlation-based models, we can already rule out certain interventions. Then, whenever you get new data, you can learn a more accurate causal model and thereby further shrink the space of interventions," Uhler explains.

This smaller search space, coupled with the acquisition function's special focus on the most informative interventions, is what makes their approach so efficient.

The researchers further improved their acquisition function using a technique known as output weighting, inspired by the study of extreme events in complex systems. This method carefully emphasizes interventions that are likely to be closer to the optimal intervention.

"Essentially, we view an optimal intervention as an 'extreme event' within the space of all possible, suboptimal interventions and use some of the ideas we have developed for these problems," Sapsis says.

Enhanced efficiency

They tested their algorithms using real biological data in a simulated cellular reprogramming experiment. For this test, they sought a genetic perturbation that would result in a desired shift in average gene expression. Their acquisition functions consistently identified better interventions than baseline methods through every step in the multi-stage experiment.

"If you cut the experiment off at any stage, ours would still be more efficient than the baselines. This means you could run fewer experiments and get the same or better results," Zhang says.

The researchers are currently working with experimentalists to apply their technique toward cellular reprogramming in the lab.

Their approach could also be applied to problems outside genomics, such as identifying optimal prices for consumer products or enabling optimal feedback control in fluid mechanics applications.

In the future, they plan to enhance their technique for optimizations beyond those that seek to match a desired mean. In addition, their method assumes that scientists already understand the causal relationships in their system, but future work could explore how to use AI to learn that information, as well.

More information: Zhang, J. et al. Active learning for optimal intervention design in causal models. *Nature Machine Intelligence* (2023). DOI: 10.1038/s42256-023-00719-0.
www.nature.com/articles/s42256-023-00719-0

(web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.