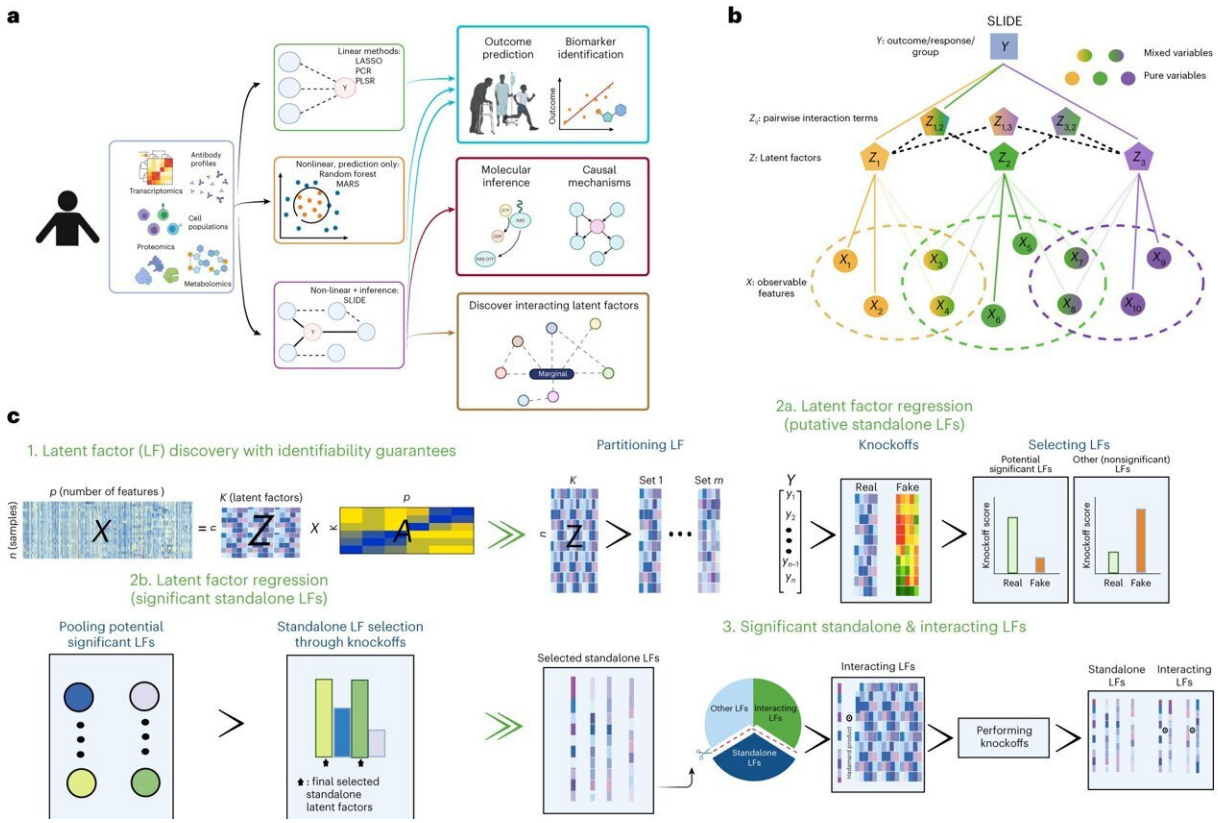
Statistical machine learning can find unknown factors that cause disease

March 27 2024, by Patricia Waldron



SLIDE—a novel interpretable machine learning method for Significant Latent Factor Interaction Discovery and Exploration. a, Schematic illustrating the vast array of datasets on which SLIDE can be applied and the key advances over existing analytical frameworks for the analyses of these datasets. b, Conceptual overview of the SLIDE algorithm. c, Schematic summarizing the implementation and different steps in SLIDE. d, Key conceptual innovations of SLIDE. e, Comparison of the predictive performance of ER, LASSO, PCR, PLSR and SLIDE on simulated datasets across a range of number of features without (left)

and with (right) interaction terms. MSE, mean squared error. f, Comparison of the predictive performance of ER, LASSO, PCR, PLSR and SLIDE on simulated datasets across a range of sample sizes without (left) and with (right) interaction terms. Credit: *Nature Methods* (2024). DOI: 10.1038/s41592-024-02175-z

A new method can now find previously unknown factors that underlie disease by using statistical machine learning to sort through mountains of complex biological data.

This flagship method, called SLIDE, successfully integrates multiple complex biological datasets and pulls out unique factors—in English, making results easy to understand—that directly or indirectly explain the data.

It may transform how we think about multi-omics data—large and varied datasets that can include detailed information on the genetics, metabolism and functions of a cell, tissue or individual, according to Cornell researchers and a Cornell Ph.D. now at the University of Pittsburgh.

Their study, "SLIDE: Significant Latent Factor Interaction Discovery and Exploration across Biological Domains," appears in *Nature Methods*.

"I love it because it is interpretable," said co-author Florentina Bunea, professor of statistics and data science in the Cornell Ann S. Bowers College of Computing and Information Science. "Essentially, we can find interpretable hidden mechanisms from measurable biological input."

The study builds on a foundation of theoretical work conducted by co-

authors including Bunea; Marten Wegkamp, professor of statistics and data science in Cornell Bowers CIS, and of mathematics in the College of Arts and Sciences; and Xin Bing, Ph.D., a former Cornell doctoral student in the field of statistics who is now at the University of Toronto.

SLIDE offers both confirmation and discovery, Bunea said, because it can corroborate previous findings and point to unknown mechanisms.

To develop this application, theoreticians at Cornell partnered with Jishnu Das, Ph.D., assistant professor of immunology at the University of Pittsburgh, a systems immunologist who studied computational biology at Cornell, where he took a stats class with Bunea.

SLIDE represents an advance over previous methods, which can only take multi-omics data profiles from samples and predict whether the samples are from healthy or diseased organisms.

"That's simply a prediction," Das said. "That's the 'what'—it doesn't get to the 'how' or the 'why.' As a biologist, I deeply care about the how and the why."

The researchers demonstrated the efficacy of SLIDE using data from 24 patients with systemic scleroderma, an autoimmune disorder that causes skin thickening and can also impair internal organs. Using data from skin biopsies that showed which genes were turned on in individual cells, researchers were able to predict the severity of the disease for each patient as well as—or better than—state-of-the-art methods.

They also identified nine hidden factors underlying the severity of the condition. Some of these factors are well-established, while others are novel, such as a previously unknown role for keratinocytes, the primary cell in the outermost layer of the skin. Additional lab experiments are already underway to confirm that the factors SLIDE identified are

indeed causing the disease symptoms.

And the paper outlines how Das's lab also used SLIDE to reproduce the locations of different types of immune cells across the lymph node in a mouse model of asthma. Similarly, in a mouse model of Type 1 diabetes, SLIDE successfully identified factors that drive the proliferation of $CD4^+$ T cells, which attack the cells in the pancreas that make insulin, resulting in the disease.

"We really believe it will be a transformative technology across disease contexts, from looking at disease severity to cellular characteristics to mechanisms of disease pathogenesis to specific cell types involved in driving these processes," Das said.

Bunea describes this collaboration between theoreticians and applied researchers as "a synergy that paid off," noting that the statistical guarantees that the hidden factors are unique and identifiable are what gives the method its power.

"The more that theory people get involved in real applications," she said, "the better it will be for all of us."

Other contributors to the paper include co-first authors Javad Rahimikollu and Hanxi Xiao from the University of Pittsburgh.

More information: Javad Rahimikollu et al, SLIDE: Significant Latent Factor Interaction Discovery and Exploration across biological domains, *Nature Methods* (2024). DOI: 10.1038/s41592-024-02175-z