Interpretable Causal Representation Learning for Biological Data in the Pathway Space

Jesus de la Fuente[1,2], **Robert Lehmann**[3], **Carlos Ruiz-Arenas**[1], **Irene Marin-Goni**[1,4], **Xabier Martinez-de-Morentin**[1], **David Gomez-Cabrero**[3], **Idoia Ochoa**[2,4], **Jesper Tegner**[3], **Vincenzo Lagani**[3,5,*] and **Mikel Hernaez**[1,6,*]

[1] CIMA, University of Navarra, IdiSNA, Pamplona, Spain.
[2] TECNUN, University of Navarra, San Sebastián, Spain.
[3] Biological and Environmental Science and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia
[4] Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic, Rochester, MN, 55905, USA
[5] Institute of Chemical Biology, Ilia State University, Tbilisi 0162, Georgia
[6] Center for Data Science (DATAI), University of Navarra, 31008, Pamplona, Spain.
[*] Corresponding authors.

Abstract

Predicting the impact of genomic and drug perturbations in cellular function is crucial for understanding gene functions and drug effects, ultimately leading to improved therapies. To this end, Causal Representation Learning (CRL) constitutes one of the most promising approaches, as it aims to identify the latent factors that causally govern biological systems, thus facilitating the prediction of the effect of unseen perturbations. Yet, current CRL methods fail in reconciling their principled latent representations with known biological processes, leading to models that are not interpretable. To address this major issue, in this work we present **SENA**-discrepancy-VAE, a model based on the recently proposed CRL method discrepancy-VAE, that produces representations where each latent factor can be interpreted as the (linear) combination of the activity of a (learned) set of biological processes. To this extent, we present an encoder, **SENA**, that efficiently compute and map biological processes' activity levels to the latent causal factors. We show that **SENA**-discrepancy-VAE achieves predictive performances on unseen combinations of interventions that are comparable with its original, non-interpretable counterpart, while inferring causal latent factors that are biologically meaningful.

AI for New Drug Modalities at NeurIPS 2024.