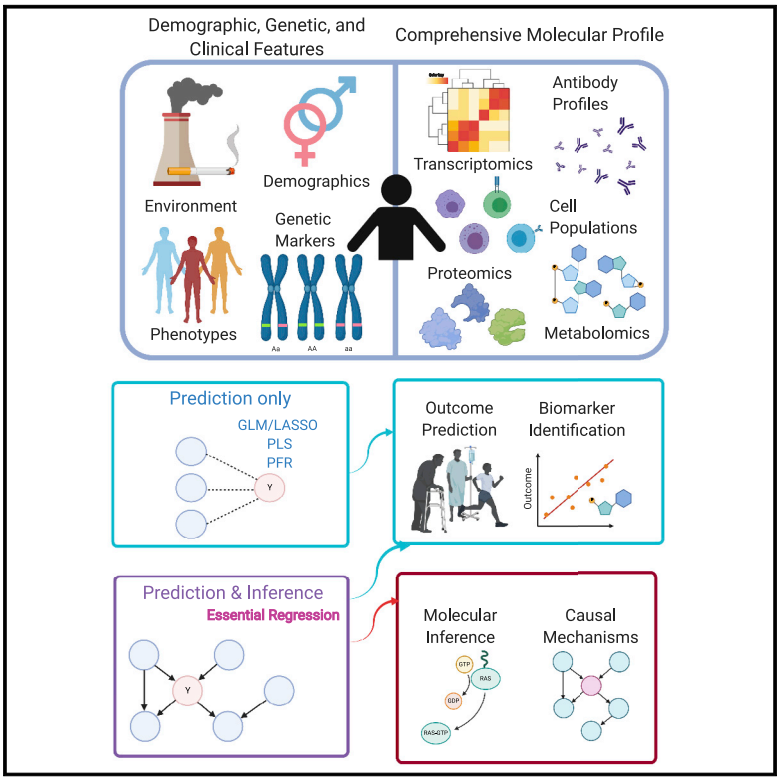
Article

Patterns

Essential Regression: A generalizable framework for inferring causal latent factors from multi-omic datasets

Graphical abstract



Highlights

- ER is a novel interpretable machine-learning method for high-dimensional multi-omic data
- ER outperforms a wide range of state-of-the-art methods in terms of prediction
- Beyond prediction, ER identifies causal latent factors of groups/outcomes of interest
- ER generated novel immunological inferences, consistent with evidence in model organisms

Authors

Xin Bing, Tyler Lovelace, Florentina Bunea, ..., Harinder Singh, Panayiotis V. Benos, Jishnu Das

Correspondence

harinder@pitt.edu (H.S.), benos@pitt.edu (P.V.B.), jishnu@pitt.edu (J.D.)

In brief

Current analytical approaches for multi-omic datasets are limited by high dimensionality, differences in data distributions, and causal inference beyond prediction. Here, we present Essential Regression (ER), a novel latent-factor-regression-based interpretable machine-learning approach that integrates high-dimensional multi-omic datasets without distributional assumptions regarding the data and identifies significant latent factors and their causal relationships with system-wide outcomes/properties of interest. ER outperforms a range of state-of-the-art methods in terms of prediction and generates novel immunological inferences, consistent with evidence in model organisms.

Bing et al., 2022, Patterns 3, 100473
May 13, 2022
https://doi.org/10.1016/j.patter.2022.100473

CellPress

Patterns

CellPress
OPEN ACCESS

Article

Essential Regression: A generalizable framework for inferring causal latent factors from multi-omic datasets

Xin Bing,[1,7] Tyler Lovelace,[2,3,7] Florentina Bunea,[1] Marten Wegkamp,[1,4] Sudhir Pai Kasturi,[5] Harinder Singh,[6,*] Panayiotis V. Benos,[2,*] and Jishnu Das[6,8,*]

[1]Department of Statistics and Data Science, Cornell University, Ithaca, NY, USA
[2]Department of Computational & Systems Biology, University of Pittsburgh, Pittsburgh, PA, USA
[3]Joint CMU-Pitt PhD Program in Computational Biology, Carnegie Mellon – University of Pittsburgh, Pittsburgh, PA, USA
[4]Department of Mathematics, Cornell University, Ithaca, NY, USA
[5]Division of Microbiology and Immunology, Yerkes National Primate Research Center, Emory University, Atlanta, GA, USA
[6]Center for Systems Immunology, Departments of Immunology and Computational & Systems Biology, University of Pittsburgh, Pittsburgh, PA, USA
[7]These authors contributed equally
[8]Lead contact
*Correspondence: harinder@pitt.edu (H.S.), benos@pitt.edu (P.V.B.), jishnu@pitt.edu (J.D.)
https://doi.org/10.1016/j.patter.2022.100473

THE BIGGER PICTURE Multi-omic technologies for deep cellular and molecular profiling from model organisms or humans have rapidly expanded. However, existing analytical approaches are constrained by the high dimensionality of these datasets, differences in data distributions, and the inability to generate causal inference beyond predictive biomarkers. To address these issues, we developed a novel interpretable machine-learning framework, Essential Regression (ER). ER integrates high-dimensional multi-omic datasets without distributional assumptions regarding the data and identifies significant latent factors and their causal relationships with system-wide outcomes/properties of interest. ER uses higher-order relationships encapsulated in the latent factors, rather than the individual observables, to home in on novel mechanistic insights. Our approach outperforms a range of state-of-the-art methods in terms of prediction and generates novel immunological inferences, consistent with evidence in model organisms.

Proof-of-Concept: Data science output has been formulated, implemented, and tested for one domain/problem

SUMMARY

High-dimensional cellular and molecular profiling of biological samples highlights the need for analytical approaches that can integrate multi-omic datasets to generate prioritized causal inferences. Current methods are limited by high dimensionality of the combined datasets, the differences in their data distributions, and their integration to infer causal relationships. Here, we present Essential Regression (ER), a novel latent-factor-regression-based interpretable machine-learning approach that addresses these problems by identifying latent factors and their likely cause-effect relationships with system-wide outcomes/properties of interest. ER can integrate many multi-omic datasets without structural or distributional assumptions regarding the data. It outperforms a range of state-of-the-art methods in terms of prediction. ER can be coupled with probabilistic graphical modeling, thereby strengthening the causal inferences. The utility of ER is demonstrated using multi-omic system immunology datasets to generate and validate novel cellular and molecular inferences in a wide range of contexts including immunosenescence and immune dysregulation.

INTRODUCTION

Over the last decade, genomic, proteomic, metabolomic, and other technologies for generating deep molecular profiles of tissues and cells from model organisms or humans have rapidly expanded.[1–4] However, the explosion in data, especially from a range of such ''omic'' technologies, has not been coupled to a proportional increase in our understanding of the underlying causal mechanisms. Existing analytical approaches have primarily focused on individual omic datasets, with relatively few attempts at integration of multi-omic datasets. In either case, we[5–9] and others[10–12] have primarily emphasized the delineation of predictive biomarkers with limited exploration of putative causal factors based on prior biological knowledge (Figure 1A). A key focus of these efforts has been to overcome the ''curse of dimensionality'' (a very large number of variables being measured in relation to a comparatively low number of samples) and the multiplicity of predictive signatures due to multi-collinear data, i.e., large correlated sets of variables. While there are several methods for reliably uncovering predictive markers from high-dimensional data, none of these analyze cause-effect relationships in relation to the outcomes/outputs of interest. This in turn has hampered efforts to undertake perturbative/translational experiments and/or clinical investigations that can test a functionally prioritized set of hypotheses generated by the large datasets.

In addition to the high dimensionality of datasets at any given scale of organization (e.g., cellular, molecular), biological systems, particularly in humans, manifest extreme complexity in terms of numbers of molecular components and their interaction rules as well as their hierarchical scales of organization, which include macromolecular complexes/condensates, organelles, cells, tissues, and organs. Each scale of organization in such a complex system has components and interaction rules that are unique to its level of organization. Thus, predicting changes in properties or behaviors of the system based on measuring components that are operating at different scales of organization represents a formidable challenge. Methods that make assumptions regarding data-generating mechanisms typically perform poorly at multi-scale integration as there are key differences in data distributions at each scale of organization.

We propose a novel framework, Essential Regression (ER), to address these key challenges and limitations of existing approaches by focusing on latent factors rather than observables in high-dimensional datasets that are significantly associated with a system-wide property or outcome that is of interest (Figure 1A). Critically, ER makes no assumptions regarding the underlying data distributions, enabling principled integration of multi-omic datasets. ER is also fundamentally different from three kinds of modern approaches. The first kind of approaches are designed specifically for multi-modal single-cell data,[13] i.e., they require single-cell data as inputs. These are constrained by structural and/or distributional requirements. ER can work on any multi-omic datasets as there are no structural assumptions regarding the data; it can even combine bulk and single-cell multi-omic datasets. The second set of approaches require prior knowledge.[14] However, ER works without the use of any priors, making it suitable across contexts even when prior knowledge is weak or unavailable. The third set of approaches[15,16] provide accurate prediction (i.e., predictive markers/correlates) from high-dimensional multi-collinear multi-omic datasets but not meaningful inference with provable statistical guarantees. ER uses regression on the latent factors rather than the observables, a novel statistical framework that comes with rigorous guarantees regarding both prediction and inference.

Overall, our analytical framework derives causal latent factors from thousands of variables from multi-omics datasets across various scales of biological organization (Figure 1A). After identifying significant latent factors, ER can be coupled with causal graphical-model analyses to examine the connectivity of these factors to the system-wide property or outcome of interest. In so doing, ER generates a high-confidence and prioritized set of latent factors comprised of known observables that are most proximal in the causal graph network to the system property/outcome of interest. We note that while causal discovery approaches have become popular over the last two decades, they have been confined to low-dimensional datasets due to the associated computational complexity. ER overcomes this fundamental conceptual limitation by first identifying latent factors from the observables (which achieves an inherent dimensionality reduction) and then identifying which latent factors are causally linked to the outcome/system-wide property of interest.

By analyzing both simulated and real-world immunological multi-omic datasets, we demonstrate that ER and the associated causal graphical-model analyses significantly outperform a wide range of state-of-the-art approaches in predicting outcomes and provide multi-scale inferences not afforded by the existing methods. The novel causal predictions are corroborated by biological findings in relevant experimental systems.

RESULTS

ER: A novel data-distribution-free statistical regression framework for inferring causal latent factors

We present ER, a novel data-distribution-free latent factor regression approach that integrates high-dimensional multi-omic datasets and identifies latent factors that are significantly associated with a system property/outcome (Figures 1B and 1C; experimental procedures; Note S1). ER is a paradigm-altering concept in regression analysis for high-dimensional datasets i.e., datasets where the number of features exceeds the number of samples. Existing regression methods use techniques including regularization (e.g., L1 regularization: least absolute selection and shrinkage operator [LASSO],[15] L1 + L2 regularization: Elastic Net), bootstrap aggregation (e.g., random forest[16]), or the incorporation of pre-specified group structures (e.g., group LASSO) to avoid overfitting. However, biomarkers/features identified using these approaches are merely predictive/correlative and may have no connection with the underlying mechanisms driving the system property/outcome of interest. ER, on the other hand, uses a two-step approach that allows for the identification of latent factors that can be used to infer causal structures underlying the system property/outcome. ER first finds latent factors in a data-dependent fashion, without the need for a pre-specified group structure. Of all latent factors, ER identifies a specific subset of latent factors that can be used to infer causal associations with the property/outcome of interest. Critically, ER makes no assumptions regarding the

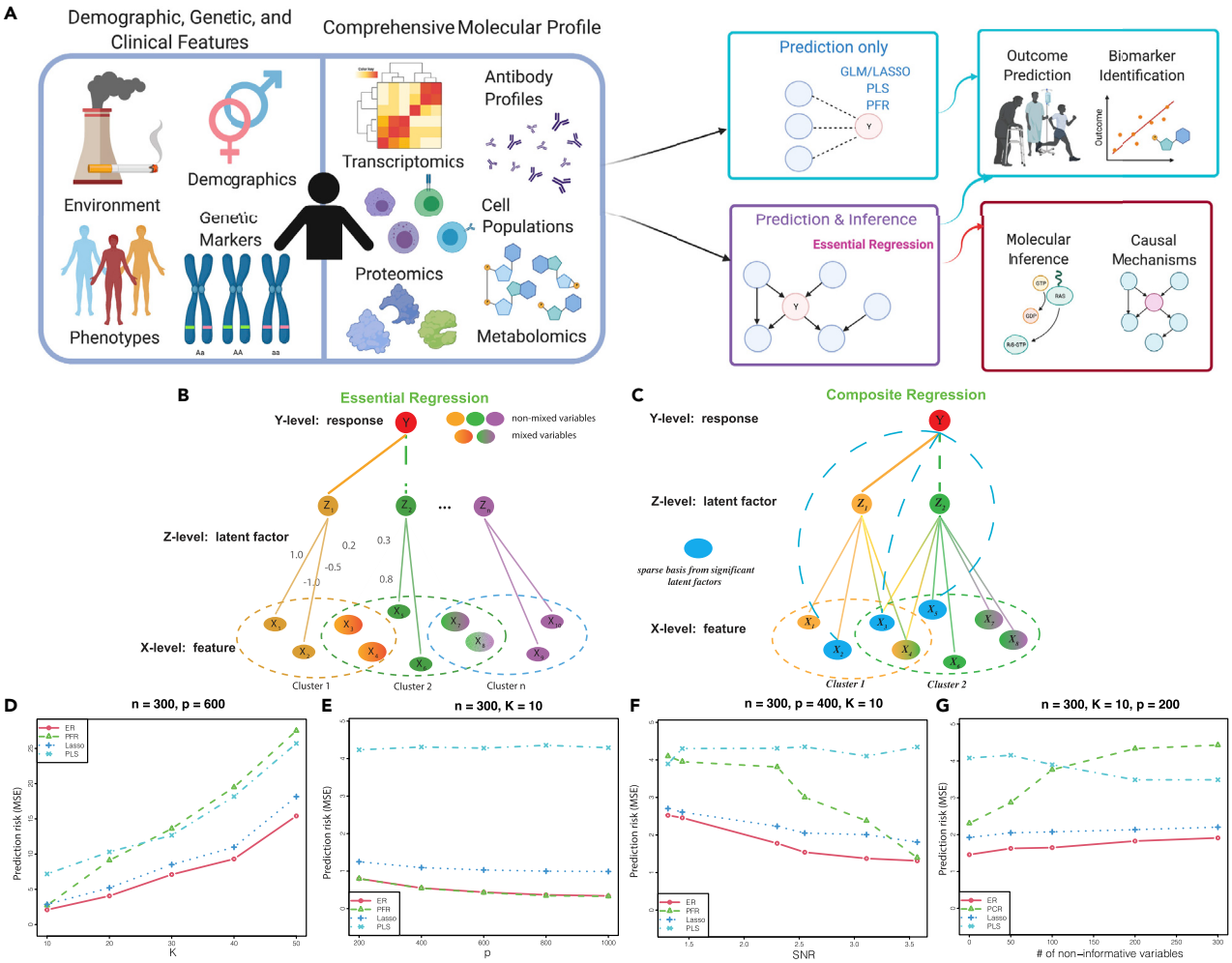


Figure 1. Essential Regression: A novel interpretable machine-learning approach to uncover causal latent factors from high-dimensional multi-omic datasets

(A) Schematic illustrating the different kinds of multi-omic datasets typically used in systems analyses and the key advantages of the methods introduced in this study over existing approaches.

(B) Schematic summarizing the steps in ER.

(C) Schematic summarizing the steps in Composite Regression.

(D–G) Comparison of the predictive performance of PLS, PFR, LASSO, and ER on simulated datasets across a range of parameter settings.

underlying data generating mechanisms and can be broadly used across multi-omic datasets (experimental procedures; Note S1).

Formally, ER is a latent-factor regression model in which the unobservable factor Z influences linearly both the response Y and the data X. Its novelty lies in the formulation that enables the latent factor Z to be meaningfully interpreted.

$$X = AZ + W$$

$$Y = \beta' Z + \varepsilon$$

Here, X is the matrix of observables and belongs to a high (p) dimensional space (dimensionality of X is $p \times n$, where n is the number of observations/samples). X is decomposed into the allocation matrix A of dimension $p \times K$, and Z is the latent-factor matrix of dimension $K \times n$, i.e., it reduces X from a p to a K dimensional space (Note S1). The matrix Z is used to regress to Y, i.e., the regression coefficients correspond to Zs. W and ε are independent error terms (Note S1). This formulation helps cluster the observables (Xs) into overlapping clusters/latent factors (Zs) in a data-dependent fashion and then identify which of the latent factors are significantly associated with and can be used to infer the outcome.

ER comes with two key provable statistical guarantees. The first step is to decompose the matrix of observables X into the latent factor matrix Z. To do this, the membership matrix A needs to be identifiable, up to a $K \times K$ signed permutation matrix. The first guarantee ensures this: we prove that the allocation matrix (A) is indeed identifiable up to a $K \times K$ signed permutation matrix under the assumption that there are at least 2 observables anchoring each latent factor (Note S1).[17] This is a reasonable

assumption as the model only requires each latent factor to be defined by two observables that are not associated with other latent factors; all other observables may or may not be associated with multiple latent factors. This allows for the identification of a group structure from the observables entirely in a data-dependent fashion without the need to incorporate any prior knowledge. The second guarantee relates to the identifiability of the regression coefficients. We also prove that the coefficient matrix is identifiable up to a signed permutation matrix (Note S1),[18] ensuring that the model can rigorously infer significant latent factors driving the outcome.

It is important to note that our current model assumes linearity at 2 different levels: (1) between the observables (*X*s) and the latent factors (*Z*s) and (2) between the outcome/response variable of interest (*Y*) and the latent factors (*Z*s). However, this does not necessarily translate into a linearity assumption between the *X*s and *Y* (it only translates into a linearity assumption when *X* and *Y* are Gaussian). Thus, the current model can incorporate non-linear relationships between *X* and *Y*, when *X* and *Y* are not Gaussian. Further, the linearity assumption between *Y* and *Z* is reasonable even for moderately large *K* (number of latent factors) as $K << p$ (p is the dimensionality of the original dataset). Further, both our algorithm and associated theoretical guarantees are still valid for a moderate and large *K*, even when $K \sim n$ (n = number of samples). Thus, ER is a first-in-class interpretable machine-learning framework that can uncover significant latent factors associated (linearly or, in many instances, non-linearly) with any system-wide property/outcome of interest.

ER outperforms state-of-the-art approaches on simulated high-dimensional datasets

We investigated the performance of ER on simulated data (Note S1), comparing it with a suite of state-of-the-art approaches including LASSO,[15] partial least squares (PLS) regression,[19] and principal components/factors regression (PFR).[20] We evaluated the performance of ER, PFR, PLS, and LASSO changes across a range of parameters for original dimensionality (p), reduced dimensionality of the dataset (i.e., *K*), and the signal-to-noise ratio (*SNR*). We varied these parameters one at a time and computed the corresponding prediction risk (mean squared error) on data not used to build the model (Note S1). We did a grid search on the relevant parameters and found that the prediction error for all four methods deteriorates as *K* increases or the *SNR* decreases (Figures 1D–1F). This indicates that prediction becomes more difficult for large *K* and a small *SNR*. On the other hand, ER, PFR, and Lasso perform better as p increases.

Among the four methods, ER systematically had the smallest prediction error in all settings, and PLS had the worst performance in most settings. Furthermore, PFR failed to accurately identify *K* and tended to a very low and sub-optimal *K* in most scenarios (Figures 1D–1F). This also indicates that, for principal-component regression approaches, detecting *K* requires a larger SNR, i.e., the other approaches are able to accurately detect *K* at lower SNRs. In a moderate SNR regime, PFR has comparable performance to ER (Figure 1D). However, as *K* increases, the advantage of ER becomes considerable, which supports the fact that PFR only has guarantees for fixed *K* (Figure 1E). Further, the performance of PFR is more sensitive to the SNR compared with the other three methods (Figure 1F). Finally, when increasing the number of uninformative variables, ER has the best performance (Figure 1G). Overall, ER worked very well for very high p, was able to accurately identify *K*, and did not have a significant reduction in performance at lower SNR regimes or with a higher number of uninformative variables (Figures 1D–1G), outperforming state-of-the-art approaches on one or more of these fronts. ER functions counterintuitively when challenged by the curse of dimensionality (i.e., having higher dimensionality is worse as it induces higher variance and can lead to overfitting). The higher dimensionality of the datasets generates more features that provide additional information, which are used by ER to predict the latent factors (*Z*) more accurately, thereby overcoming the curse of dimensionality.

Extension of ER as composite regression enables uncovering of observables, within significant latent factors, that underlie outcomes

While the significant latent factors uncovered by ER provide insights into the interplay of the different observables driving outcome, in some contexts their complexity can prove challenging. In these instances, smaller sets of observables underlying outcome are desirable. Currently, regularization is widely used to identify a sparse set of observables (biomarkers). However, regularization-based approaches such as LASSO or Elastic Net uncover predictive biomarkers that may simply be correlative. Given that ER identifies latent factors significantly driving outcome, we sought to develop an approach to identify a sparse set of observables from the significant latent factors identified by ER (Figure 1C). Using L1-regularization on the significant latent factors identified by ER allows us to identify a sparse set of observables, within these factors, tied to outcome. We term this ER-derivative-approach Composite Regression (CR) (Figure 1C). As the sparse set of observables delineated by CR are selected from those that lie within the significant latent factors, unlike LASSO-based biomarkers, these are no longer simply predictive but capture causal relationships that can be used to infer the underlying mechanistic basis of outcome. Together, ER and CR provide a highly prioritized set of significant latent factors and associated observables, which can be used both for inference of underlying cellular/molecular mechanisms as well as corresponding biomarkers.

Inferring causal factors underlying immunosenescence in a vaccine response

A recent study comprehensively profiled cellular and molecular responses induced by the shingles Zostavax vaccine in a cohort comprising both younger adults and elderly individuals.[21] The high-dimensional multi-omic analysis included immune-cell frequencies and phenotypes, as well as transcriptomic, metabolomic, cytokine, and antibody analyses. The vaccine induced robust antigen-specific antibody titers as well as $CD4^+$, but not $CD8^+$, T cell responses.[21] Using a multi-scale, multifactorial response network, the authors identified associations between transcriptomic, metabolomic, cellular phenotypic, and cytokine datasets that pointed to immune and metabolic correlates of vaccine immunity.[21] Interestingly, differences in the quality of the vaccine-induced responses by age were also noted.[21] We hypothesized that a method based on latent factors rather than measurables

would improve the delineation of components that underlie the quality and magnitude of the vaccine-induced responses. If so, then such a method would be able to leverage the differences in vaccine-induced responses and accurately predict age as the system-wide property of interest. The latent factors identified in this manner could then provide insights into the cellular and molecular basis of age-induced immunosenescence manifested by diminished responses to the Zostavax vaccine.

To explore the above formulation of immunosenescence as a predictor of age, we first applied a suite of state-of-the-art approaches, LASSO, PLS, and PFR, on the entire spectrum of multi-omic vaccine-induced responses (including transcriptomic, metabolomic, cytokine, antibody, and cellular phenotypic data) to predict age (Figure 2A). As most individuals in the cohort were in 2 distinct age groups, adults under 40 and elderly people over 60, we first sought to explore the performance of LASSO, PLS, and PFR in predicting the two age groups as binary categorical variables, i.e., younger adults and elderly people. The predictive performance of all methods was evaluated in a stringent leave-one-out cross-validation (LOOCV) framework (experimental procedures). We have previously demonstrated that on such multi-omic datasets, cross validation is a gold standard to evaluate model performance with data held out.[5,6,8] In an LOOCV framework, we found that PFR had no predictive power (area under the curve [AUC] <0.5), while LASSO and PLS had weak predictive power, in predicting age as a categorical variable (Figure 2B, AUCs = 0.63 and 0.60, respectively). The receiver operating characteristic (ROC) curve for LASSO had an interesting shape. It attained a true positive rate of ~0.4 at a false positive rate of ~0.15, but beyond that it was essentially no better than random (Figure 2B). This observation is consistent with the observation that differences in an age-associated multiscale multifactorial response network (MMRN) were driven by only a subset of elderly vaccinees.[21] Thus, a purely predictive modeling approach like LASSO can leverage these relatively straightforward differences to accurately predict age for a subset of the vaccinees but fails to predict age for others. We then compared these methods with the performance of ER and CR. In a matched, LOOCV framework, ER and CR were very accurate at predicting age (Figure 2B, AUCs = 0.79 and 0.77, respectively, $p < 0.01$).

We then coupled ER to causal-inference analyses on the ER-identified significant latent factors using directed graphical models.[22] Directed acyclic graphs (DAGs) are sometimes referred to as causal graphs because under certain assumptions the learned DAGs from observational data (Markov equivalence classes) asymptotically represent the true data-generating causal graph. Although these algorithms have shown considerable success in analyzing many biological processes and biomedical problems,[23–27] including biomarker selection and classification,[28–30] scalability limits the datasets to which they can be applied.[31,32] Here, we use the causal-learning algorithm for mixed data, CausalMGM,[23,33] only on the significant latent factors delineated by ER to overcome the scale limitation. By applying CausalMGM only on the significant latent factors, we greatly reduce the dimensionality of the input dataset while preserving the information of individual (correlated) variables in the latent factors. Thus, CausER (CausalMGM on the significant latent factors from ER) prioritizes further within the significant latent factors (experimental procedures; Note S1) by virtue of their direct connections to the outcome in the graphical model. Furthermore, it predicts potential cause-effect relationships between the latent factors and the property/outcome of interest, which leads to hypotheses generation, while CausER was the best predictor of age as a categorical variable (AUC = 0.86, $p < 0.01$). Together, these results demonstrate that while LASSO, PLS, and PFR fail to accurately predict age from Zostavax-induced vaccine responses, ER, CR, and CausER can overcome this challenging problem by leveraging non-trivial differences in latent factors comprised of discrete sets of measurables.

Next, we evaluated whether these methods could predict actual age as a continuous variable beyond the categorical classifiers of younger adults and elderly individuals. As before, performance was measured in a rigorous cross-validation framework (experimental procedures). Using the vaccine-induced responses, PFR was not at all predictive of age (Figure 2C, Pearson $r = -0.71$; Figure S1, Spearman $r = -0.82$). LASSO and PLS had poor performance in predicting age as a continuous variable (Figure 2C, Pearson $r = 0.29$ and 0.13, respectively; Figure S1, Spearman $r = 0.25$ and 0.09, respectively). In fact, the predictive powers of PLS and PFR were not significantly different from a negative control model built on permuted data (Figure 2C). However, both ER and CR were significantly predictive of age as a continuous variable (Pearson $r = 0.48$ for both, Spearman $r = 0.44$ and 0.49, respectively, $p < 0.01$; Figures 2C and S1), and as in the previous instance, CausER had the best performance in predicting age as a continuous variable (Pearson $r = 0.61$, Spearman $r = 0.59$, $p < 0.01$; Figures 2C and S1). Together, these results demonstrate that while state-of-the-art methods including LASSO, PLS, and PFR fail to predict age either as a categorical or a continuous variable, all three of the new approaches that are based on latent factors, ER, CR, and CausER, are able to do so reasonably accurately based on the multi-omic profiles of vaccine-induced responses.

We next explored the likely causal relationships among the latent factors that lead to age-induced immunosenescence and diminished responses to the Zostavax vaccine. CausalMGM was used to construct a causal graph with all latent factors identified in the latent-model-identification step of ER (Figure 2D). Notably, the majority of significant latent factors identified by ER were seen to be proximal to the outcome variable (age) in the causal graph. Importantly, all 4 latent factors in the Markov blanket generated by CausalMGM were also identified as significant by ER (Figure 2D). Overall, the significant latent factors revealed by ER had significantly lower network distances (i.e., they had stronger cause-effect relationships) from age compared with the non-significant latent factors (Figure 2E, $p < 0.05$). These results demonstrate that the cause-effect relationships identified by ER are validated by CausalMGM. Importantly, while CausER hits are identified via the sequential application of ER and CausalMGM, respectively, the order is critical, with ER being the key first step. Without the two-stage dimensionality reduction (first from observables to latent factors and then the identification of significant latent factors) afforded by ER, running CausalMGM or other allied causal graphical models on the initial set of observables would be computationally intractable.

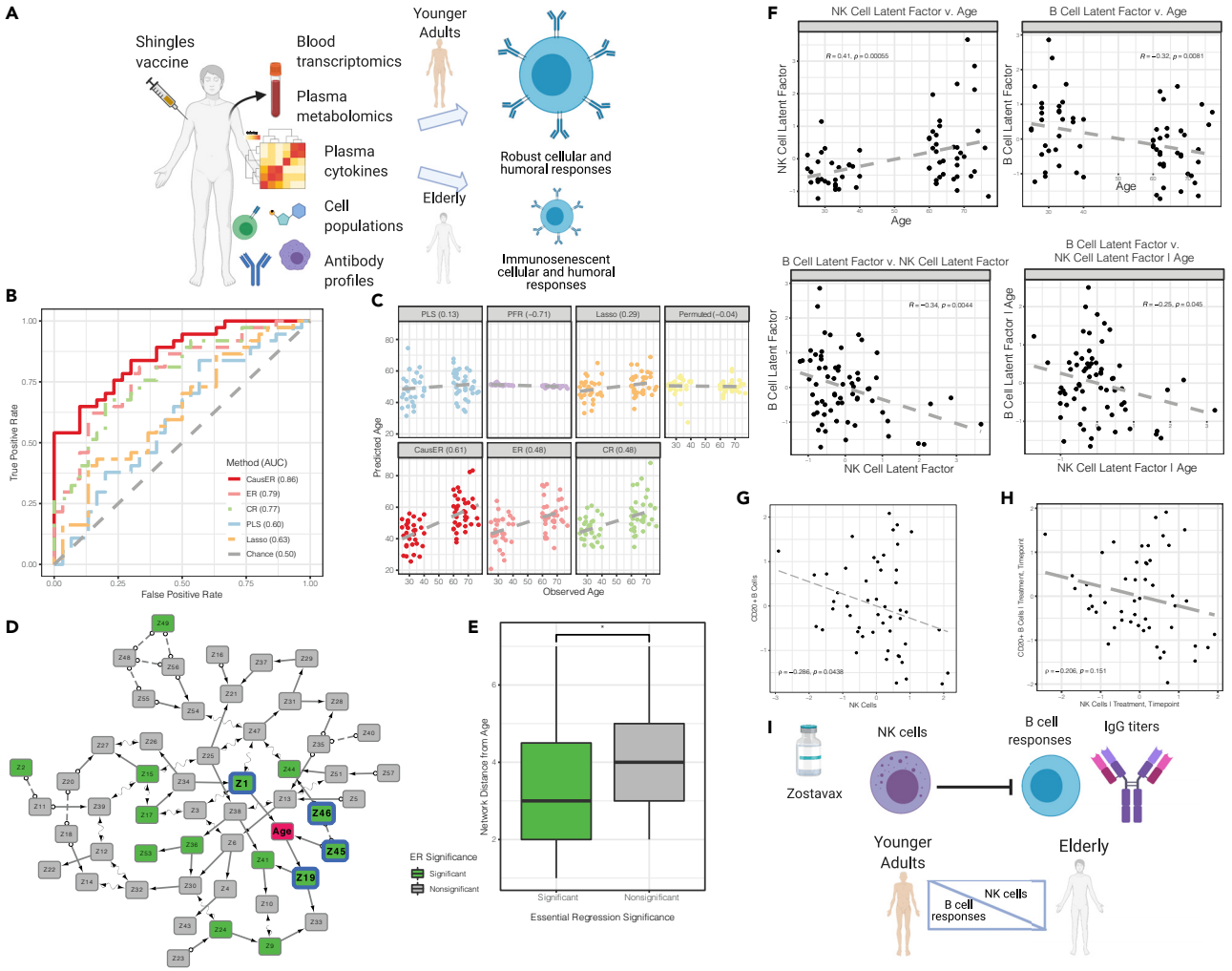


Figure 2. Identifying causal signatures of age-induced immunosenescent responses to the Zostavax vaccine

(A) Schematic summarizing the input data and the problem of interest.

(B) ROC curves for the different methods at discriminating between elderly people and younger adults in an LOOCV framework.

(C) Pearson correlations of the different methods at predicting age as a continuous variable, as measured in an LOOCV cross-validation framework.

(D) CausalMGM on all Zs identified by ER. The Markov blanket is highlighted with a blue border and bolder fonts. A directed edge X → Y indicates X is a cause of Y, while a bidirected edge X ← → Y indicates the presence of a latent confounder that is a common cause of X and Y. A partially oriented edge X o→ Y indicates that Y is not a cause of X but that either X or a latent confounder causes Y. Unoriented edge indicates directionality could not be inferred for that edge.

(E) Network distances in the causal graph generated by CausalMGM of the significant and non-significant Zs identified by ER from the outcome variable of interest. p value calculated using a Mann-Whitney U test

(I) Mechanistic insights obtained from ER.

(F) Correlations involving the NK cell latent factor, B cell latent factor, and age. Top panels show correlations between the NK cell latent factor and age (top left), and the B cell latent factor and age (top right). Bottom panels show correlations between the NK cell latent factor and the B cell latent factor without correcting for age (bottom left) and after correcting for age (bottom right).

(G) Correlations between NK cells and B cells in the context of vaccination against SARS-CoV2 in a NHP model.

(H) Correlations between NK cells and B cells in the context of vaccination against SARS-CoV2 in a NHP model, after correcting for treatment (vaccination arm) and timepoint.

The prioritized CausER hits (Figure 2D), i.e., significant latent factors identified by ER that are also in the Markov blanket of the outcome variable (age) in the causal graph generated by CausalMGM, comprised antigen-specific immunoglobulin G (IgG) titers (Z1), a metabolic module (Z19), and B (Z46) and natural killer (NK; Z45) cell frequencies. CausER provides both prioritized cause-effect relationships and directions of these relationships. While the latter relates to mathematical conditional-independence relationships (experimental procedures), the former provides prioritized mechanistic insights. Notably, the discovery and labeling of causal latent factors are completely unbiased and not based on prior knowledge. These significant latent factors are those that were identified as significant by ER and in the Markov blanket of outcome; neither step used any prior knowledge.

However, to evaluate the quality of these discoveries, we examined the uncovered latent factors in light of previously

elucidated bases of immunosenescence. The lowering of titers with age is expected and has been previously reported,[21] so this corresponds to a recapitulation of known relationships. However, CausER also revealed a novel cause-effect relationship between altered B and NK cell numbers and immunosenescence. To further dissect the nature of this relationship, we examined correlations between NK cells, B cells, and age. We found that NK cells significantly increased, while the numbers of B cells significantly decreased, with age (Figure 2F). More interestingly, there was a significant negative correlation between NK and B cells (Figure 2F), and the correlation remained significant even after correcting for age (Figure 2F).

Notably, these causal inferences are supported by perturbation experiments involving biologically relevant organisms. Our findings relate to a previously described mechanistic linkage between NK cells and a weaker germinal center (GC) response in a murine model[34] in the context of vaccination with a model antigen (NP-KLH). NK cells can inhibit CD4 T cell responses, including those of T follicular helper cells, in a perforin-dependent manner; this leads to a weaker GC response and diminished antibody titers and affinity maturation.[34,35] Furthermore, in the context of vaccination against severe acute respiratory syndrome coronavirus 2 (SARS-CoV2) in a non-human primate (NHP) model, we leveraged cell-subset-frequency data from a recent study[36] to examine the relationship between NK and B cells. We found a significant negative relationship between NK and B cells spanning multiple time points and vaccination arms corresponding to different adjuvants (Figure 2G). These relationships remained unaltered even after correcting for time point and vaccination arm using a linear model (Figure 2H). Together, these results demonstrate that a novel relationship uncovered solely by ER and CausER, without the use of any prior knowledge, from a human-systems vaccinology study have strong support in vaccination studies both in mice and NHPs. Notably, these studies use different antigens and adjuvants, suggesting that the uncovered novel relationship between NK and B cells is highly robust, and can be broadly extrapolated across vaccination strategies. Our results suggest a novel basis of human immunosenescence in the context of vaccine responses (Figure 2I). This discovery is especially striking as ER converged on this mechanism without the use of any prior knowledge.

Analyzing latent factors potentially reflective of trained immunity in a vaccine response

To test whether ER is applicable to datasets generated using alternate technological platforms, we applied it to analyze the temporal dynamics of transcriptional responses (microarray data) induced by the malaria RTS,S vaccine.[37] RTS,S has a standard regimen of 3 doses separated by a month and is currently the most advanced malaria-vaccine candidate that has consistently demonstrated 40%–80% protective efficacy in malaria-naïve individuals in controlled human challenge studies.[5] There has been intense interest over the last decade in uncovering molecular signatures induced by the RTS,S vaccine and corresponding correlates of protection.[5,38,39] In a controlled human-infection setting, differential expression of immunoproteasome genes was identified as a pre-challenge correlate of protection.[37] After the third dose, as expected, there was a striking but transitory shift in inflammatory gene expression followed a convergence of the majority of gene signatures back to pre-vaccination levels within 2 weeks after the third dose.[37] We reasoned that aspects of trained immunity induced by the vaccine may be reflected in the transcriptomic signatures that do not converge after 2 weeks. Thus, a sensitive method such as ER would be able to discriminate between expression profiles at the following time points, pre-vaccination (G1), the day after the third dose (G2), and 14 days after the third dose (G3) (Figure 3A), and reveal candidate genes and molecular pathways that could contribute to trained immunity. In this instance, the use of a microarray dataset also afforded the opportunity to explore how ER performs with noisier but nevertheless valuable datasets generated using older technologies.

As before, the ability of the different methods to discriminate between G1, G2, and G3 transcriptional profiles was measured in a rigorous cross-validation framework (experimental procedures). We found that there were significant differences in the ability of the different methods to discriminate between the three kinds of expression profiles, with ER and CausER (CausalMGM on the significant latent factors from ER) having the best performance, significantly better than the other methods ($p < 0.01$, Figures 3B and 3C). Next, we chose to focus on the ability of the different methods to specifically distinguish the G3 profile from the other two (Figure 3D) or just the G1 profile (Figure 3E). This constituted the most "difficult" discrimination as there are broad differences in the expression profiles between the pre- (G1) and 24-hour-post-vaccination (G2) time points, but most of these differences disappear by 14 days (G3).[37] Consistent with expectations, in this binary-classification setting, there was wide variability in the performance of the methods to specifically discriminate the G3 time point from the G1 and G2 time points. While PFR and PLS performed poorly, CausER, ER, and LASSO had significantly better performances, with CausER being the best-performing method ($p < 0.01$; Figures 3D and 3E). In terms of correctly classifying just the true G3 profiles as G3, PLS and PFR had poor performances while CausER had the best performance, significantly better than other methods ($p < 0.01$, Figure 3F).

Next, we focused on the CausER hits, i.e., the significant latent factors from ER in the Markov blanket of the outcome variable (Figure 3G). Genes comprising these latent factors were seen to be differentially expressed between the G1 and G3 samples (Figures 3H and 3I). Our results suggest that beyond the initial divergence of immunoproteasome genes, there is a sustained divergence (2 weeks post-vaccination) of genes involved in immune-metabolic processes. These results complement recent findings that suggest that targeting immunometabolism is a promising direction in modulating trained immunity.[40] While a vaccine induces a rapid initial divergence in inflammatory signatures reflecting the activation of innate immune cells and their engagement with adaptive B and T cells, it may also induce alterations in the innate immune compartment that are discernible at later time points and contribute to a distinct form of immune memory.[40]

Elucidating markers of latent and active tuberculosis (Tb)

To explore whether ER and CausER can predict clinically important outcomes, we applied these approaches to a dataset of high-dimensional antibody profiles for patients with latent and active Tb[41] (Figure 4A). The high-dimensional antibody-omic dataset used a modern antibody-omic platform[5,6,8,41] to quantify

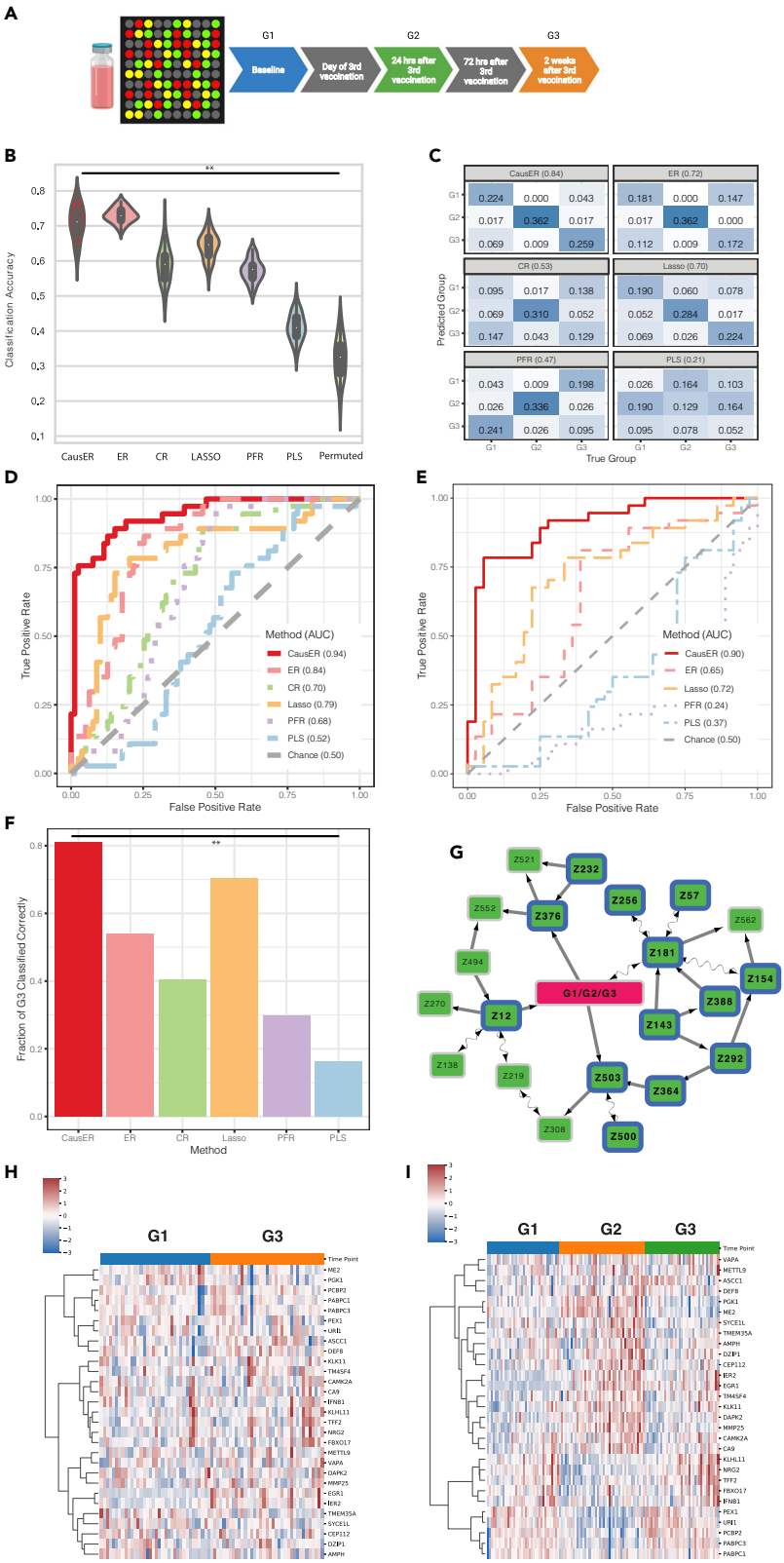


Figure 3. Identifying differences in vaccine-induced transcriptomic profiles over time

(A) Schematic summarizing the input data and the problem of interest.

(B) Ternary classification accuracy of the different methods at discriminating among G1, G2, and G3 in a replicated *k*-fold cross-validation framework.

(C) Confusion matrix summarizing the performance of the different methods at discriminating among G1, G2, and G3 in an LOOCV framework.

(D) ROC curves for the different methods at discriminating between G3 and G1 and G2 combined in an LOOCV framework.

(E) ROC curves for the different methods at discriminating between G3 and G1 in an LOOCV framework.

(F) Fraction of true G3 correctly classified as G3 (as measured in an LOOCV framework).

(G) CausER graph i.e., CausalMGM on the significant Zs from ER. The Markov blanket is highlighted with a blue border and bolder fonts. A directed edge X → Y indicates X is a cause of Y, while a bidirected edge X ←→ Y indicates the presence of a latent confounder that is a common cause of X and Y. A partially oriented edge X o→ Y indicates that Y is not a cause of X but that either X or a latent confounder causes Y. Unoriented edge indicates directionality could not be inferred for that edge.

(H) Heatmap of genes in CausER hits (significant Zs in the Markov blanket) for G1 and G3 samples.

(I) Heatmap of genes in CausER hits (significant Zs in the Markov blanket) for G1, G2, and G3 samples.

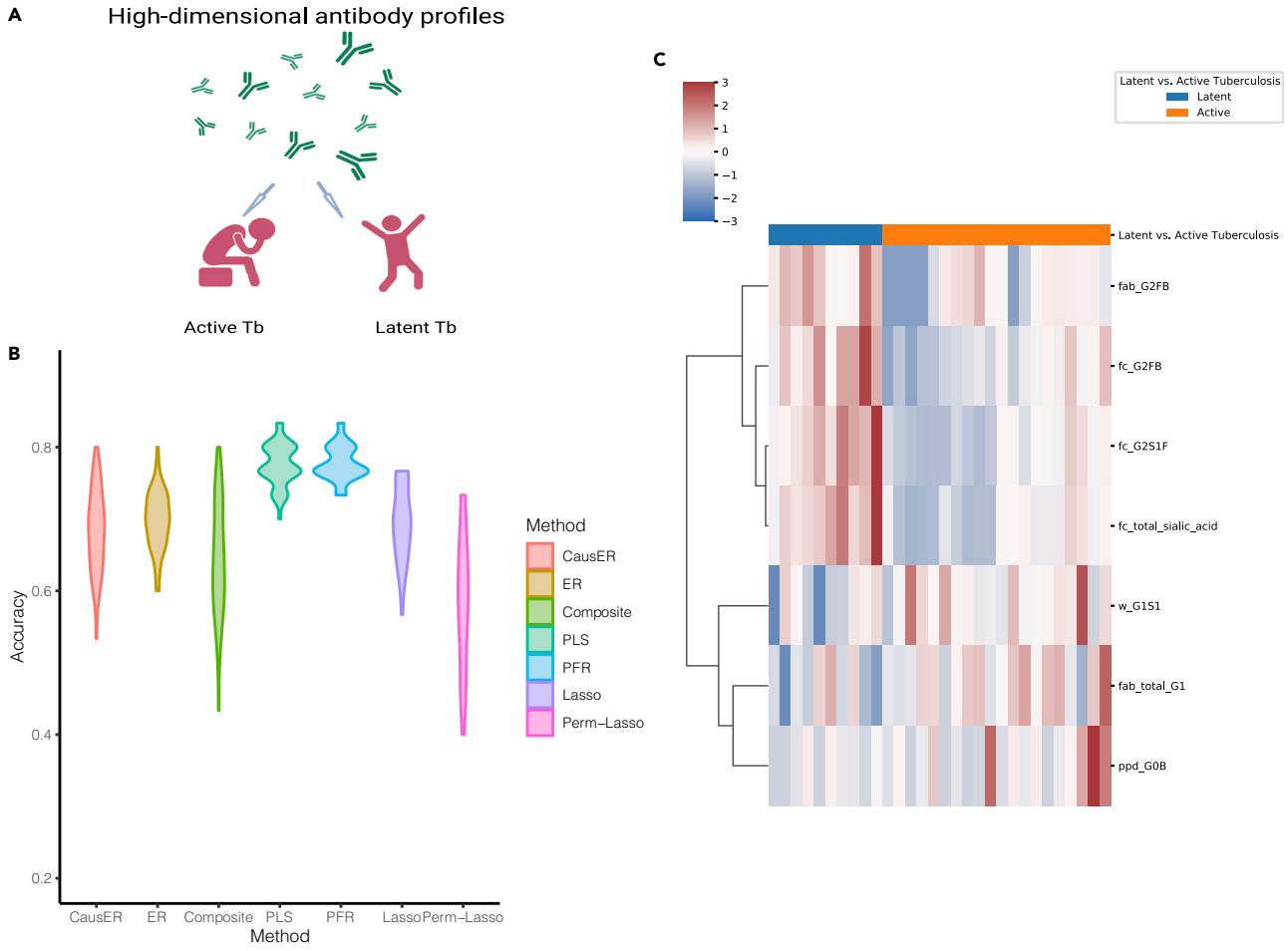


Figure 4. Elucidating markers of latent and active tuberculosis (Tb)

(A) Schematic summarizing the input data and the problem of interest.

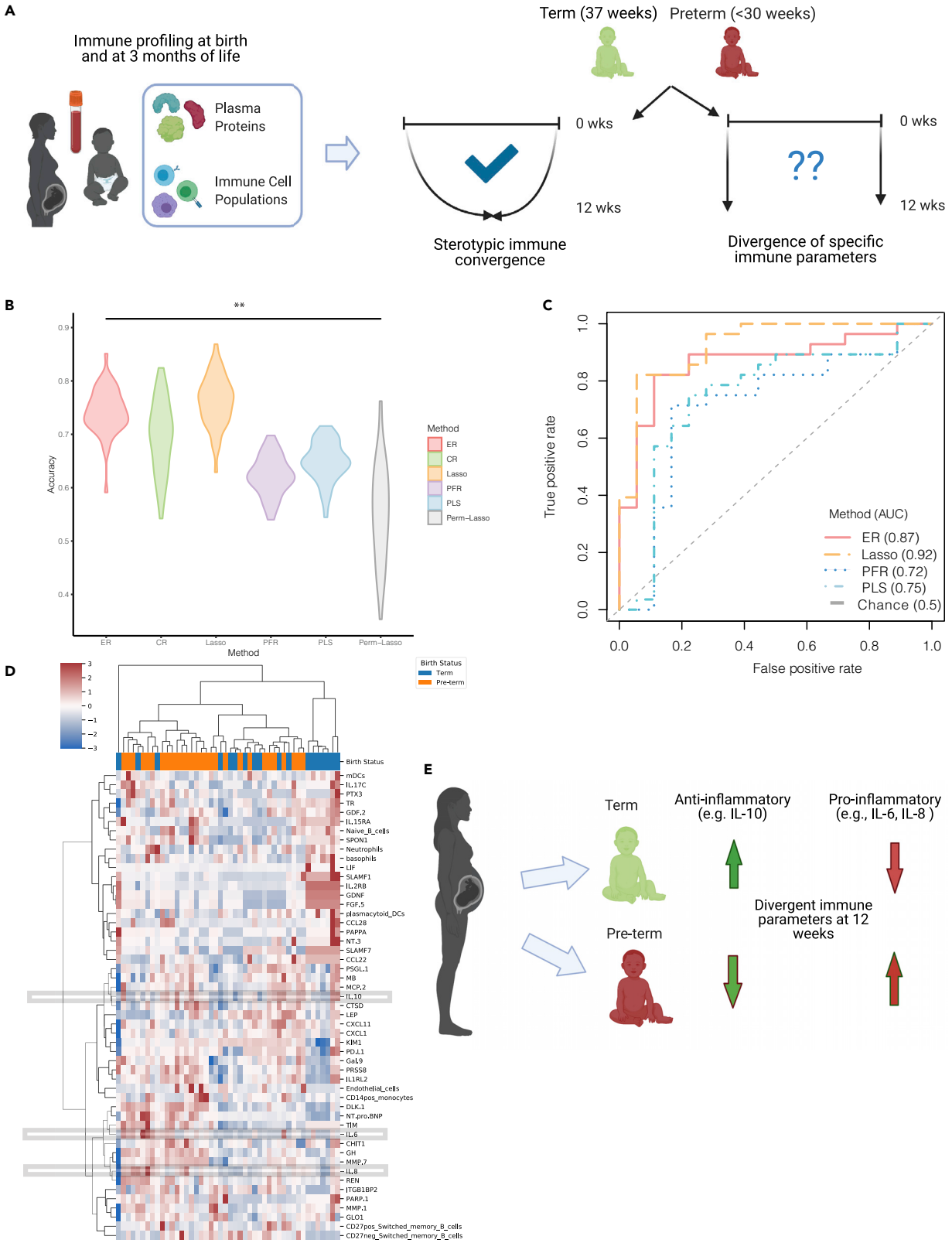
(B) Classification accuracy of the different methods at discriminating between latent and active Tb, measured in a replicated *k*-fold cross-validation framework.

(C) Heatmap of features in the single CausER hit.

functional and biophysical properties of a polyclonal pool of antigen-specific antibodies. Each of these properties has its own inherent distribution so this was an appropriate test of the ability of ER and CausER to integrate multi-modal datasets for clinical outcome prediction. CausER and ER along with PLS, PFR, and LASSO were able to accurately discriminate between latent and active Tb patients using the antibody-omic profiles (Figure 4B). Notably, only one latent factor was identified as significant by ER and in the Markov blanket of outcome, i.e., this latent factor was the sole CausER hit. It consisted of specific glycosylation profiles (Figure 4C), and the majority of these glycosylation-based biomarkers were in perfect agreement with our previous study.[41] These analyses demonstrate that ER and CausER are able to accurately predict clinically important outcomes.

Uncovering latent factors that distinguish immune-system states of term and pre-term infants

Finally, we focused on a multi-omic longitudinal cohort that analyzed immune-cell populations and plasma proteins in 100 newborn children during their first 3 months of life[42] (Figure 5A). Striking differences were observed in immune parameters between pre-term and term children at birth. However, the immune trajectories appeared to achieve a stereotypic convergence within the first 3 months of life[42] (Figure 5A). We hypothesized that ER might be able to uncover latent factors that distinguish immune-system states of term and pre-term infants after 3 months of life and therefore reveal features that could impact later life (Figure 5A). As expected, based on the striking differences at birth between term and pre-term children, all methods (LASSO, PLS, PFR, ER, CR, and CausER) were able to discriminate between these 2 groups using immune parameters measured in the first week of life (Figure S2). All model performances were measured in a rigorous cross-validation framework (experimental procedures). However, given the stereotypic convergence in the first 3 months (12 weeks) of life,[42] we found that PLS and PFR were unable to accurately discriminate between term and pre-term children using immune parameters measured at 12 weeks of life (Figures 5B and 5C). However, LASSO was able to accurately distinguish between term and pre-term births using the 12-week profiles (Figures 5B and 5C),



(legend on next page)

suggesting that despite broad convergence, a small subset of immune parameters still remain different in term and pre-term infants at 3 months of life. More importantly, ER and CR were able to accurately discriminate between term and pre-term births using immune profiles at 3 months of life, significantly better than other methods (Figures 5B and 5C, $p < 0.01$). ER identified only 2 significant latent factors, and based on CausalMGM analyses, one of these 2 significant latent factors was in the Markov blanket, i.e., for this dataset, this single latent factor was the sole CausER hit (Figure 5D).

We visualized the immune-cell populations and plasma proteins in this hit (Figure 5D). These profiles had clearly remained divergent even at 3 months of life (Figure 5D) despite the broad stereotypic convergence of most other immune parameters. At 3 months of life, term infants had an anti-inflammatory milieu including high interleukin (IL)-10 while pre-term infants had a pro-inflammatory milieu including elevated IL-6 and IL-8 (Figure 5D). These findings agree with a previous study that IL-10 is highly expressed in the uterus and placenta and has a key role in controlling inflammation-induced pre-term labor in a murine model.[43] Furthermore, regulatory B cells are a key source of IL-10 and appear to be important in sustaining pregnancy until term.[44–46] It is also known that modulation of pro- versus anti-inflammatory environments by relevant cytokines and chemokines at the maternal-fetal interface (decidua) is a critical component of the bifurcation between term and pre-term births.[44] Thus, our analyses of immune-system states of term and pre-term infants at 3 months of life revealed that pre-term infants had a pro-inflammatory state while term infants had an anti-inflammatory state (Figure 5E). These findings could have long-term implications for the health of pre-term infants.

DISCUSSION

Over the last two decades, while there have been rapid advances in high-throughput experimental technologies to generate deep molecular profiles, computational analyses of these high-dimensional datasets have primarily focused on biomarker discovery.[47] This is because rigorous statistical approaches for analyzing high-dimensional datasets, such as regularized regression and bootstrap-aggregated classification, are focused on uncovering predictive biomarkers, which may simply be correlative surrogates of outcome or system-wide property but are unrelated to the underlying causal factors. Incorrect extrapolation of insights derived from biomarker-based approaches can lead to perturbation experiments with low success. Alternatively, efforts to move beyond biomarkers to mechanistic insights often use biological priors, which may be incomplete or suffer from sampling/study biases.[48] Furthermore, while there have been advances in causal modeling,[49] existing approaches are difficult to apply to high-dimensional datasets due to the computational intractability of applying these approaches on[22] and the multi-collinearity of the data. The methods presented in this article address this fundamental limitation in systems biology. ER is a first-in-class machine-learning method that can both handle high-dimensional multi-omic datasets with co-linear variables and prioritize cause-effect relationships between the input features and the outcome of interest. Our framework is also complementary to modern approaches that combine multi-omic datasets with prior knowledge to uncover causal relationships.[14] ER generates mechanistic hypotheses solely based on latent factors identified from multi-omic data without the incorporation of any prior knowledge. It is thus applicable in contexts where prior knowledge is weak or unavailable and is not limited by the nature and quality of available prior knowledge. ER is compatible with all existing batch correction/normalization approaches as it makes no assumptions regarding data-generating mechanisms. However, data need to be appropriately normalized/batch corrected before being used as inputs to ER. Further, ER is also able to handle complex replicate structures. Biological and technical replicates may be pre-processed using a suitable context-specific approach; ER does not impose any restrictions on/is robust to how replicates are handled (which depends entirely on the underlying biological context/question). ER works downstream of these methods to integrate appropriately pre-processed/normalized multi-omic datasets and uncover causal latent factors underlying groups/outcomes of interest.

Importantly, ER is fundamentally different from classical factor regression models used exclusively for prediction. In those models, one first seeks a low-dimensional factor $Z = XV$ constructed via some projection matrix V. Although, Z can then be used to regress to Y, this framework can only be used for prediction and not inference as Z is not uniquely identifiable and this makes inference on the regression coefficients impossible. However, in the ER framework, the latent factors (Zs) and the corresponding linear coefficients (between Y and Zs) are uniquely identifiable, making the inference problem well-posed. Thus, our framework addresses a key limitation of classical factor regression models where the recovered factors have ambiguous meaning. However, the unique identifiability of the latent factors in the unsupervised step of ER makes inference meaningful. Thus, we cannot simply replace it with other modern clustering approaches with no guarantees regarding identifiability. The identifiability criterion tied to the guarantees regarding inference make ER a first-in-class interpretable latent-factor regression framework for high-dimensional multi-omic datasets.

Our framework pushes the envelope on multiple key challenges in systems biology. First, it establishes a rigorous framework with provable statistical guarantees that explores a large space of higher-order relationships from high-dimensional features and uncovers latent factors tied to the outcome variable via directed cause-effect relationships. Second, unlike existing

Figure 5. Uncovering specific immune parameters from term and pre-term infants that do not achieve stereotypic convergence

(A) Schematic summarizing the input data and the problem of interest.

(B) Classification accuracy of the different methods at discriminating between term and pre-term births using immune profiles at 3 months after birth, measured in a replicated *k*-fold cross-validation framework.

(C) ROC curves for the different methods at discriminating between term and pre-term births as measured in an LOOCV framework.

(D) Heatmap of features (plasma proteins and immune cells) in the single hit (significant Z identified by ER in the Markov blanket of outcome).

(E) Mechanistic insights obtained from ER.

causal-reasoning approaches that are constrained by the size of the input data, ER can be applied to modern high-dimensional datasets. The time complexities of the different steps are essentially quadratic and not exponential like some other causal-reasoning approaches. Third, ER makes no assumptions regarding data-generating mechanisms, and ER can integrate multi-omic datasets to capture the interplay across a plethora of biological processes at multiple scales of organization of the system. Fourth, ER is able to home in on one or a few causal latent factors of outcome comprising a small number of observable features from thousands of input features, many of which are completely uninformative. Finally, ER converges on these causal latent factors without the use of any prior knowledge; however, we find that the uncovered factors include both previously elucidated and novel mechanistic bases. The ability of ER to converge on meaningful biological insights without any prior knowledge makes it applicable in the broadest sense even in contexts where there are weak or no priors.

An important elaboration of our framework is the sequential use of two orthogonal methods for statistical inference, ER and causal graphical modeling. These methods have different theoretical bases and assumptions, and yet the ER hits are validated by CausalMGM, underscoring the robustness of our approach. The order is critical, with ER being the key first step offering a two-stage dimensionality reduction: first from observables to latent factors, and then the identification of significant latent factors. Without these two steps, the application of causal graphical models on the initial set of observables would be computationally intractable due to the high dimensionality of the dataset. Thus, ER solves a long-standing limitation with causal graphical-modeling approaches and enables, for the first time, causal inference on high-dimensional data. ER also has polynomial time complexity that makes it efficient and scalable for extremely large datasets. For all the datasets analyzed in this study, it resulted in runtimes of core minutes for each cross-validation replicate, which translates into tens to hundreds of core hours after accounting for cross-validation replicates. The datasets included tens to hundreds of samples and up to 10^3–10^4 features/sample. So, this all suggests that ER is extremely efficient with modern multi-omic datasets.

ER has a number of limitations. One relates to the constraint each latent factor is anchored by at least 2 pure variables (i.e., variables that belong to only that and no other latent factor). However, this is a reasonable assumption as most observables are allowed to be mixed, i.e., they can belong to one or more latent factors, and each latent factor only requires 2 pure variables to anchor it. Also, in some instances, the linearity assumption between Y and Z could be restrictive. For example, when the number of latent factors is small, this restrictive assumption could be overcome by including high-order terms of Z to predict Y. It is also possible to extend the current framework to a more general setting where Y and Z follow generalized linear models with any appropriate link function, such as the logistic and probit functions. While it is relatively straightforward to incorporate suitable link functions in the setting of prediction, achieving theoretical guarantees for the inference of the coefficients of Z needs more careful theoretical analyses.

The coupling of causal graphical models to ER and the inference of causality from observational data also has some assumptions. First, it is assumed that the structure of the cause-effect relationships of all variables in the dataset form a DAG. Next, the causal Markov assumption requires that the Markov condition for DAGs holds for the causal graph. Finally, the causal faithfulness assumption states that the conditional-independence relationships in the dataset are faithful to the causal graph. A distribution is faithful to its causal DAG when there are no additional conditional-independence relationships that are not entailed by the Markov condition of the DAG. Importantly, while some algorithms also require the assumption of causal sufficiency, which states that there are no unobserved confounders of the variables in the dataset, the fast causal inference (FCI) algorithm used here does not have this constraint. Further, for full identification of the causal graph, the assumptions of the conditional-independence test, in this case linearity, must be met, and the sample size must be asymptotically large. Thus, the inference of true causality is constrained by these assumptions, which may not always hold. However, importantly, these assumptions are tied to the causal graphical-modeling framework. ER (without coupling to CausalMGM) can be used to identify significant latent factors, with only very minimal assumptions, as described above. Thus, while true causality may, in some instances, be difficult to infer from observational data, the significant latent factors identified by ER provide inference into generative processes beyond just prediction.

Here, we applied ER to diverse contexts. First, we applied ER on simulated datasets and demonstrated that it performed better than LASSO, PLS, and PFR across a range of parameter settings. Next, we utilized ER on two recent human systems-immunology studies that had generated high-dimensional multi-omic profiles. Using ER, we were able to address key questions that had not been the focus of the original studies, in part because of limitations of methods used. Such questions could now be addressed by the methodological advances of ER over state-of-the-art approaches. We demonstrated that ER significantly outperforms PFR and PLS across contexts and either outperforms or matches LASSO in terms of predictive performance. While we used three examples to illustrate the superior performance of ER, these methods come with broad theoretical guarantees to outperform PLS, PFR, and LASSO across contexts (experimental procedures; Figure 1). Furthermore, while the existing methods simply identify correlates, without using any prior knowledge, ER provides mechanistic insights. Some of these outcomes are consistent with previous mechanistic experiments while others are novel. ER can also be used for noisier and older datasets not generated using state-of-the-art methods. Our findings have broad implications across domains in systems biology and are likely to transform both computational workflows used to analyze multi-omic datasets and downstream experiments designed based on the insights gleaned via these analyses.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Requests for data and code used for the study should be directed to and will be fulfilled by the lead contact, Jishnu Das (jishnu@pitt.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Detailed code, associated datasets, and documentation for ER, CR, and CausER are available at https://github.com/jishnu-lab/ER. A corresponding stable release can be accessed at https://doi.org/10.5281/zenodo.6178063.

Any queries regarding the code or data should be directed to the lead contact, Jishnu Das (jishnu@pitt.edu).

Theoretical underpinnings of ER

We provide brief descriptions of the methods, associated tuning parameters, cross-validation strategies, and data pre-processing in this section. Additional theoretical details are included in Note S1.

Processing of systems-immunology datasets

For the dataset of multi-omic responses to the Zostavax vaccine, we included the following multi-scale measurements of immune state: IgG titers, blood transcriptional modules, metabolic clusters, $CD4^+$ T cell populations, T follicular helper (TFH) cell populations, flow-cytometry cell populations, cytokine profiles, and IFN T cells. We used subject age as the response variable for $n = 72$ subjects. We excluded features that had missing values for more than a half of subjects. We also excluded 5 subjects that had no observed features. The remaining datasets were merged via the unique IDs of subjects. The final dataset contains $p = 1{,}721$ features of $n = 67$ subjects.

For the transcriptomic (microarray) dataset pre- and post-malaria vaccination, we had $n = 116$ samples with 22,277 probes. We filtered out ambiguous probes (i.e., those that could map to multiple genes) and then averaged technical replicates (multiple probes/gene) with the limma package in R. The final dataset comprised 116 samples and $p = 12{,}424$ genes. Y is a categorical variable with 3 levels corresponding to three time points.

For the dataset of high-dimensional antibody profiles, we had $n = 30$ subjects (20 latent Tb, 10 active Tb) with $p > 100$ features/subject. The features included titers, Fc effector functions and whole, Fab, and Fc glycan profiles (independent of antigen) as well PPD- and Ag85-specific titers and glycan profiles.

For the dataset of term and pre-term infants, we included all available immune parameters as features and only removed clinical metadata (such as "gender," "mode of delivery," "family," etc.). The final dataset we used has $n = 183$ samples and $p = 282$ features with 56 samples from week 1 and 46 samples from week 12. The response is binary, either "control" (representing term) or "pre-term" (representing pre-term). We used the 5-NN to impute the missing values.

Cross validation

Two cross-validation techniques were used to assess the predictive performance of the different methods: (1) replicated 10-fold cross validation and 2) LOOCV. (1) To assess the accuracy of the classifiers for the term/pre-term immune profile, 50 replicates of nested 10-fold cross validation were performed. For each replicate, we independently ran each of the methods and assessed the predictive accuracy. For ER, the latent factors were learned on each fold and each replicate, and the regression and final latent-factor selection were repeated. For CausER, a causal model was learned over the latent factors selected as significant by ER for each fold and replicate. The average cross-validation accuracy across the 10-folds was calculated for each of the 50 replicates. (2) For the datasets, we also performed LOOCV to assess the accuracy of each method. In LOOCV, each sample in the dataset was held out as the predictive models were trained on the remaining samples, and then the held-out sample was predicted with the trained models. Assessment of model performance (AUC) was done with the set of predictions for the left-out values.

ER

The first step in ER is the estimation of all latent factors. The identification of latent factors is unsupervised. This is done based on the empirical sample covariance matrix using a three-step procedure. The first step involves the identification of latent-variable structure using the sample covariance matrix. A key component of this step is the identification of K (reduced dimensionality) from p (original dimensionality). The second step involves inference of the clusters: each cluster (latent factor) is anchored by at least 2 pure variables. Variables that are associated with multiple clusters are designated mixed variables. The third step involves determination of the overall allocation matrix based on the cluster assignments in the earlier step. Formal descriptions of all 3 steps are provided in Note S1, Section 2.

After the identification of Zs, the regression coefficients linking the Zs to Y are estimated using a theoretical framework we recently established for estimation in latent-factor regression models.[50] This is the supervised part of the ER algorithm. A detailed description of the estimation procedure is provided in Note S1, Section 2.

CR

CR utilizes a 2-step procedure. First, it uses ER to identify significant Zs as described above. Then, it uses LASSO on the Xs associated only with the significant Zs to identify a sparse basis for the system-wide property/outcome of interest. For LASSO on the significant Zs identified by ER, lambda is tuned using k-fold cross validation. The lambda tuning is specific to a given fold for a given replicate and utilizes only the fold-specific training data.

ER coupled to CausalMGM

We implemented CausalMGM as previously described[22] on all Zs for the Zostavax dataset and only the significant Zs identified by ER for the term/pre-term, malaria, and Tb datasets. Briefly, when constructing the causal model, we first learned an undirected graphical model with MGM[51]/GLASSO.[52] The optimal regularization parameters were selected based on graph stability using StEPS[33]/StARS.[53] The resulting undirected graph was then used as an initial graph for performing causal inference with the FCI algorithm. To build a predictor of the outcome variable, the Markov blanket was used. The Markov blanket was defined as the set of variables that, when conditioned on, make the response variable independent of every other variable in the dataset according to the structure of the causal graph. For a DAG, this comprises the parents, children, and spouses (other parents of the children) of the response variable.

Implementation of LASSO, PLS, and PFR

LASSO was implemented using glmnet in R with parameter tuning done in a manner analogous to that described above for ER and CR. If no feature was selected by LASSO in a specific fold for a given replicate, we randomly selected 5 features (only for that fold in that replicate) and used an ordinary-least-squares estimator. Thus, the feature selection in each case is specific to each fold for a given replicate; this is the most stringent and unbiased way to evaluate model performance. PLS was implemented using the plsr function in R with the number of components selected by the default function selectNcomp. For PFR, which regresses Y on the first K principal components of X, the number of principal components K is selected based on the ratios of non-decreasing eigenvalues of X'^*X/n using previously established criteria.[54]

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2022.100473.

ACKNOWLEDGMENTS

The authors would like to thank Mirko Paiardini and Maria Elena Bottazzi for their support. This study was partially supported by NIH grants DP2AI164325 to J.D., U01HL137159, R01HL140963, R01HL159805, and R01HL157879 to P.V.B., U01AI141990 to H.S., and F31LM013966 to T.L.; NSF grants DMS-1712709 and DMS-2015195 to F.B. and M.W.; and DoD grant W81XWH2110864 to J.D. H.S. also acknowledges support from the UPMC ITTC fund. S.P.K. acknowledges support from the Yerkes Pilot Research Pilot Program (part of the Yerkes NPRC Base Grant, P51-OD011132). Several images were created with BioRender.com.

AUTHOR CONTRIBUTIONS

J.D. designed the study and oversaw all aspects of it. X.B., F.B., and M.W. jointly conceived the theoretical basis of the ER framework. J.D. and X.B. jointly conceived the application of the ER and CR frameworks to real data.

J.D., P.V.B., and H.S. jointly conceived the CausER framework. X.B. and T.L. implemented the ER, CR, and CausER frameworks and carried out all computational analyses. S.P.K. provided data. J.D., P.V.B., and H.S. interpreted the results. J.D., H.S., and P.V.B. wrote the main text. X.B., T.L., F.B., and M.W. wrote the supplementary methods including formal proofs.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 28, 2021
Revised: September 17, 2021
Accepted: March 1, 2022
Published: March 24, 2022

REFERENCES

1. Hagan, T., and Pulendran, B. (2018). Will systems biology deliver its promise and contribute to the development of new or improved vaccines? From data to understanding through systems biology. Cold Spring Harb. Perspect. Biol. *10*, a028894. https://doi.org/10.1101/cshperspect.a028894.
2. Pulendran, B., Li, S., and Nakaya, H.I. (2010). Systems vaccinology. Immunity *33*, 516–529. https://doi.org/10.1016/j.immuni.2010.10.006.
3. Davis, M.M., Tato, C.M., and Furman, D. (2017). Systems immunology: just getting started. Nat. Immunol. *18*, 725–732. https://doi.org/10.1038/ni.3768.
4. Villani, A.C., Sarkizova, S., and Hacohen, N. (2018). Systems immunology: learning the rules of the immune system. Annu. Rev. Immunol. *36*, 813–842. https://doi.org/10.1146/annurev-immunol-042617-053035.
5. Suscovich, T.J., Fallon, J.K., Das, J., Demas, A.R., Crain, J., Linde, C.H., Michell, A., Natarajan, H., Arevalo, C., Broge, T., et al. (2020). Mapping functional humoral correlates of protection against malaria challenge following RTS,S/AS01 vaccination. Sci. Transl. Med. *12*, eabb4757. https://doi.org/10.1126/scitranslmed.abb4757.
6. Das, J., Devadhasan, A., Linde, C., Broge, T., Sassic, J., Mangano, M., O'Keefe, S., Suscovich, T., Streeck, H., Irrinki, A., et al. (2020). Mining for humoral correlates of HIV control and latent reservoir size. PLoS Pathog. *16*, e1008868. https://doi.org/10.1371/journal.ppat.1008868.
7. Goetghebuer, T., Smolen, K.K., Adler, C., Das, J., McBride, T., Smits, G., Lecomte, S., Haelterman, E., Barlow, P., Piedra, P.A., et al. (2019). Initiation of antiretroviral therapy before pregnancy reduces the risk of infection-related hospitalization in human immunodeficiency virus-exposed uninfected infants born in a high-income country. Clin. Infect Dis. *68*, 1193–1203. https://doi.org/10.1093/cid/ciy673.
8. Ackerman, M.E., Das, J., Pittala, S., Broge, T., Linde, C., Suscovich, T.J., Brown, E.P., Bradley, T., Natarajan, H., Lin, S., et al. (2018). Route of immunization defines multiple mechanisms of vaccine-mediated protection against SIV. Nat. Med. *24*, 1590–1598. https://doi.org/10.1038/s41591-018-0161-0.
9. Sadanand, S., Das, J., Chung, A.W., Schoen, M.K., Lane, S., Suscovich, T.J., Streeck, H., Smith, D.M., Little, S.J., Lauffenburger, D.A., et al. (2018). Temporal variation in HIV-specific IgG subclass antibodies during acute infection differentiates spontaneous controllers from chronic progressors. AIDS *32*, 443–450. https://doi.org/10.1097/QAD.0000000000001716.
10. Vafaee, F., Diakos, C., Kirschner, M.B., Reid, G., Michael, M.Z., Horvath, L.G., Alinejad-Rokny, H., Cheng, Z.J., Kuncic, Z., and Clarke, S. (2018). A data-driven, knowledge-based approach to biomarker discovery: application to circulating microRNA markers of colorectal cancer prognosis. NPJ Syst. Biol. Appl. *4*, 20. https://doi.org/10.1038/s41540-018-0056-1.
11. Li, S., Rouphael, N., Duraisingham, S., Romero-Steiner, S., Presnell, S., Davis, C., Schmidt, D.S., Johnson, S.E., Milton, A., Rajam, G., et al. (2014). Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. Nat. Immunol. *15*, 195–204. https://doi.org/10.1038/ni.2789.
12. Nakaya, H.I., Wrammert, J., Lee, E.K., Racioppi, L., Marie-Kunze, S., Haining, W.N., Means, A.R., Kasturi, S.P., Khan, N., Li, G.M., et al. (2011). Systems biology of vaccination for seasonal influenza in humans. Nat. Immunol. *12*, 786–795. https://doi.org/10.1038/ni.2067.
13. Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J.C., and Stegle, O. (2020). MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. *21*, 111. https://doi.org/10.1186/s13059-020-02015-1.
14. Dugourd, A., Kuppe, C., Sciacovelli, M., Gjerga, E., Gabor, A., Emdal, K.B., Vieira, V., Bekker-Jensen, D.B., Kranz, J., Bindels, E.M.J., et al. (2021). Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses. Mol. Syst. Biol. *17*, e9730. https://doi.org/10.15252/msb.20209730.
15. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B (Methodological) *58*, 267–288.
16. Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32. https://doi.org/10.1023/A:1010933404324.
17. Bing, X., Bunea, F., Ning, Y., and Wegkamp, M. (2020). Adaptive estimation in structured factor models with applications to overlapping clustering. Ann. Stat. *48*, 2055–2081, 2027.
18. Bing, X., Bunea, F., and Wegkamp, M. (2019). Inference in latent factor regression with clusterable features. Preprint at arXiv. 1905.12696. https://doi.org/10.48550/arXiv.1905.12696.
19. Boulesteix, A.L., and Strimmer, K. (2007). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Brief Bioinform. *8*, 32–44. https://doi.org/10.1093/bib/bbl016.
20. Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006). Prediction by supervised principal components. J. Am. Stat. Assoc. *101*, 119–137. https://doi.org/10.1198/016214505000000628.
21. Li, S., Sullivan, N.L., Rouphael, N., Yu, T., Banton, S., Maddur, M.S., McCausland, M., Chiu, C., Canniff, J., Dubey, S., et al. (2017). Metabolic phenotypes of response to vaccination in humans. Cell *169*, 862–877.e17. https://doi.org/10.1016/j.cell.2017.04.026.
22. Ge, X., Raghu, V.K., Chrysanthis, P.K., and Benos, P.V. (2020). CausalMGM: an interactive web-based causal discovery tool. Nucleic Acids Res. *48*, W597–W602. https://doi.org/10.1093/nar/gkaa350.
23. Sedgewick, A.J., Buschur, K., Shi, I., Ramsey, J.D., Raghu, V.K., Manatakis, D.V., Zhang, Y., Bon, J., Chandra, D., Karoleski, C., et al. (2019). Mixed graphical models for integrative causal analysis with application to chronic lung disease diagnosis and prognosis. Bioinformatics *35*, 1204–1212. https://doi.org/10.1093/bioinformatics/bty769.
24. Schadt, E.E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., Sieberts, S.K., Monks, S., Reitman, M., Zhang, C., et al. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. Nat. Genet. *37*, 710–717. https://doi.org/10.1038/ng1589.
25. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A., and Nolan, G.P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. Science *308*, 523–529.
26. Manatakis, D.V., Raghu, V.K., and Benos, P.V. (2018). piMGM: incorporating multi-source priors in mixed graphical models for learning disease networks. Bioinformatics *34*, i848–i856. https://doi.org/10.1093/bioinformatics/bty591.
27. Kitsios, G.D., Fitch, A., Manatakis, D.V., Rapport, S.F., Li, K., Qin, S., Huwe, J., Zhang, Y., Doi, Y., Evankovich, J., et al. (2018). Respiratory microbiome profiling for etiologic diagnosis of pneumonia in mechanically ventilated patients. Front. Microbiol. *9*, 1413. https://doi.org/10.3389/fmicb.2018.01413.
28. Abecassis, I., Sedgewick, A.J., Romkes, M., Buch, S., Nukui, T., Kapetanaki, M.G., Vogt, A., Kirkwood, J.M., Benos, P.V., and Tawbi, H. (2019). PARP1 rs1805407 increases sensitivity to PARP1 inhibitors in cancer cells suggesting an improved therapeutic strategy. Sci. Rep. *9*, 3309. https://doi.org/10.1038/s41598-019-39542-2.

29. Raghu, V.K., Zhao, W., Pu, J., Leader, J.K., Wang, R., Herman, J., Yuan, J.M., Benos, P.V., and Wilson, D.O. (2019). Feasibility of lung cancer prediction from low-dose CT scan and smoking factors using causal models. Thorax *74*, 643–649. https://doi.org/10.1136/thoraxjnl-2018-212638.

30. Raghu, V.K., Beckwitt, C.H., Warita, K., Wells, A., Benos, P.V., and Oltvai, Z.N. (2018). Biomarker identification for statin sensitivity of cancer cell lines. Biochem. Biophys. Res. Commun. *495*, 659–665. https://doi.org/10.1016/j.bbrc.2017.11.065.

31. Raghu, V.K., Ramsey, J.D., Morris, A., Manatakis, D.V., Sprites, P., Chrysanthis, P.K., Glymour, C., and Benos, P.V. (2018). Comparison of strategies for scalable causal discovery of latent variable models from mixed data. Int. J. Data Sci. Anal. *6*, 33–45. https://doi.org/10.1007/s41060-018-0104-3.

32. Raghu, V.K., Poon, A., and Benos, P.V. (2018). Evaluation of causal structure learning methods on mixed data types. Proceedings of 2018 ACM SIGKDD Workshop on Causal Disocvery (PMLR).

33. Sedgewick, A.J., Shi, I., Donovan, R.M., and Benos, P.V. (2016). Learning mixed graphical models with separate sparsity parameters and stability-based model selection. BMC Bioinformatics *17* (*Suppl 5*), 175. https://doi.org/10.1186/s12859-016-1039-0.

34. Rydyznski, C., Daniels, K.A., Karmele, E.P., Brooks, T.R., Mahl, S.E., Moran, M.T., Li, C., Sutiwisesak, R., Welsh, R.M., and Waggoner, S.N. (2015). Generation of cellular immune memory and B-cell immunity is impaired by natural killer cells. Nat. Commun. *6*, 6375. https://doi.org/10.1038/ncomms7375.

35. Rydyznski, C.E., Cranert, S.A., Zhou, J.Q., Xu, H., Kleinstein, S.H., Singh, H., and Waggoner, S.N. (2018). Affinity maturation is impaired by natural killer cell suppression of germinal centers. Cell Rep. *24*, 3367–3373.e4. https://doi.org/10.1016/j.celrep.2018.08.075.

36. Pino, M., Abid, T., Pereira Ribeiro, S., Edara, V.V., Floyd, K., Smith, J.C., Latif, M.B., Pacheco-Sanchez, G., Dutta, D., Wang, S., et al. (2021). A yeast expressed RBD-based SARS-CoV-2 vaccine formulated with 3M-052-alum adjuvant promotes protective efficacy in non-human primates. Sci. Immunol. *6*, eabh3634. https://doi.org/10.1126/sciimmunol.abh3634.

37. Vahey, M.T., Wang, Z., Kester, K.E., Cummings, J., Heppner, D.G., Jr., Nau, M.E., Ofori-Anyinam, O., Cohen, J., Coche, T., Ballou, W.R., and Ockenhouse, C.F. (2010). Expression of genes associated with immunoproteasome processing of major histocompatibility complex peptides is indicative of protection with adjuvanted RTS,S malaria vaccine. J. Infect Dis. *201*, 580–589. https://doi.org/10.1086/650310.

38. Kazmin, D., Nakaya, H.I., Lee, E.K., Johnson, M.J., van der Most, R., van den Berg, R.A., Ballou, W.R., Jongert, E., Wille-Reece, U., Ockenhouse, C., et al. (2017). Systems analysis of protective immune responses to RTS,S malaria vaccination in humans. Proc. Natl. Acad. Sci. U S A *114*, 2425–2430. https://doi.org/10.1073/pnas.1621489114.

39. Neafsey, D.E., Juraska, M., Bedford, T., Benkeser, D., Valim, C., Griggs, A., Lievens, M., Abdulla, S., Adjei, S., Agbenyega, T., et al. (2015). Genetic diversity and protective efficacy of the RTS,S/AS01 malaria vaccine. N. Engl. J. Med. *373*, 2025–2037. https://doi.org/10.1056/NEJMoa1505819.

40. Arts, R.J., Joosten, L.A., and Netea, M.G. (2016). Immunometabolic circuits in trained immunity. Semin. Immunol. *28*, 425–430. https://doi.org/10.1016/j.smim.2016.09.002.

41. Lu, L.L., Das, J., Grace, P.S., Fortune, S.M., Restrepo, B.I., and Alter, G. (2020). Antibody Fc glycosylation discriminates between latent and active tuberculosis. J. Infect Dis. *222*, 2093–2102. https://doi.org/10.1093/infdis/jiz643.

42. Olin, A., Henckel, E., Chen, Y., Lakshmikanth, T., Pou, C., Mikes, J., Gustafsson, A., Bernhardsson, A.K., Zhang, C., Bohlin, K., and Brodin, P. (2018). Stereotypic immune system development in newborn children. Cell *174*, 1277–1292.e14. https://doi.org/10.1016/j.cell.2018.06.045.

43. Robertson, S.A., Skinner, R.J., and Care, A.S. (2006). Essential role for IL-10 in resistance to lipopolysaccharide-induced preterm labor in mice. J. Immunol. *177*, 4888–4896. https://doi.org/10.4049/jimmunol.177.7.4888.

44. Gomez-Lopez, N., StLouis, D., Lehr, M.A., Sanchez-Rodriguez, E.N., and Arenas-Hernandez, M. (2014). Immune cells in term and preterm labor. Cell Mol. Immunol. *11*, 571–581. https://doi.org/10.1038/cmi.2014.46.

45. Rolle, L., Memarzadeh Tehran, M., Morell-Garcia, A., Raeva, Y., Schumacher, A., Hartig, R., Costa, S.D., Jensen, F., and Zenclussen, A.C. (2013). Cutting edge: IL-10-producing regulatory B cells in early human pregnancy. Am. J. Reprod. Immunol. *70*, 448–453. https://doi.org/10.1111/aji.12157.

46. Jensen, F., Muzzio, D., Soldati, R., Fest, S., and Zenclussen, A.C. (2013). Regulatory B10 cells restore pregnancy tolerance in a mouse model. Biol. Reprod. *89*, 90. https://doi.org/10.1095/biolreprod.113.110791.

47. Libbrecht, M.W., and Noble, W.S. (2015). Machine learning applications in genetics and genomics. Nat. Rev. Genet. *16*, 321–332. https://doi.org/10.1038/nrg3920.

48. Cusick, M.E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A.R., Simonis, N., Rual, J.F., Borick, H., Braun, P., Dreze, M., et al. (2009). Literature-curated protein interaction datasets. Nat. Methods *6*, 39–46. https://doi.org/10.1038/nmeth.1284.

49. Pearl, J. (2010). An introduction to causal inference. Int. J. Biostat. *6*. Article 7. https://doi.org/10.2202/1557-4679.1203.

50. Bing, X., Bunea, F., Strimas-Mackey, S., and Wegkamp, M. (2021). Prediction in latent factor regression: adaptive PCR, interpolating predictors and beyond. J. Mach. Learn. Res. *22*, 1–50. https://www.jmlr.org/papers/volume22/20-768/20-768.pdf.

51. Lee, J.D., and Hastie, T.J. (2015). Learning the structure of mixed graphical models. J. Comput. Graph Stat. *24*, 230–253. https://doi.org/10.1080/10618600.2014.900500.

52. Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. Biostatistics *9*, 432–441. https://doi.org/10.1093/biostatistics/kxm045.

53. Liu, H., Roeder, K., and Wasserman, L. (2010). Stability approach to regularization selection (StARS) for high dimensional graphical models. Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2 (Curran Associates Inc.).

54. Lam, C., and Yao, Q. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. Ann. Stat. *40*, 694–726, 633.

Patterns, Volume *3*

Supplemental information

Essential Regression: A generalizable framework for inferring causal latent factors from multi-omic datasets

Xin Bing, Tyler Lovelace, Florentina Bunea, Marten Wegkamp, Sudhir Pai Kasturi, Harinder Singh, Panayiotis V. Benos, and Jishnu Das

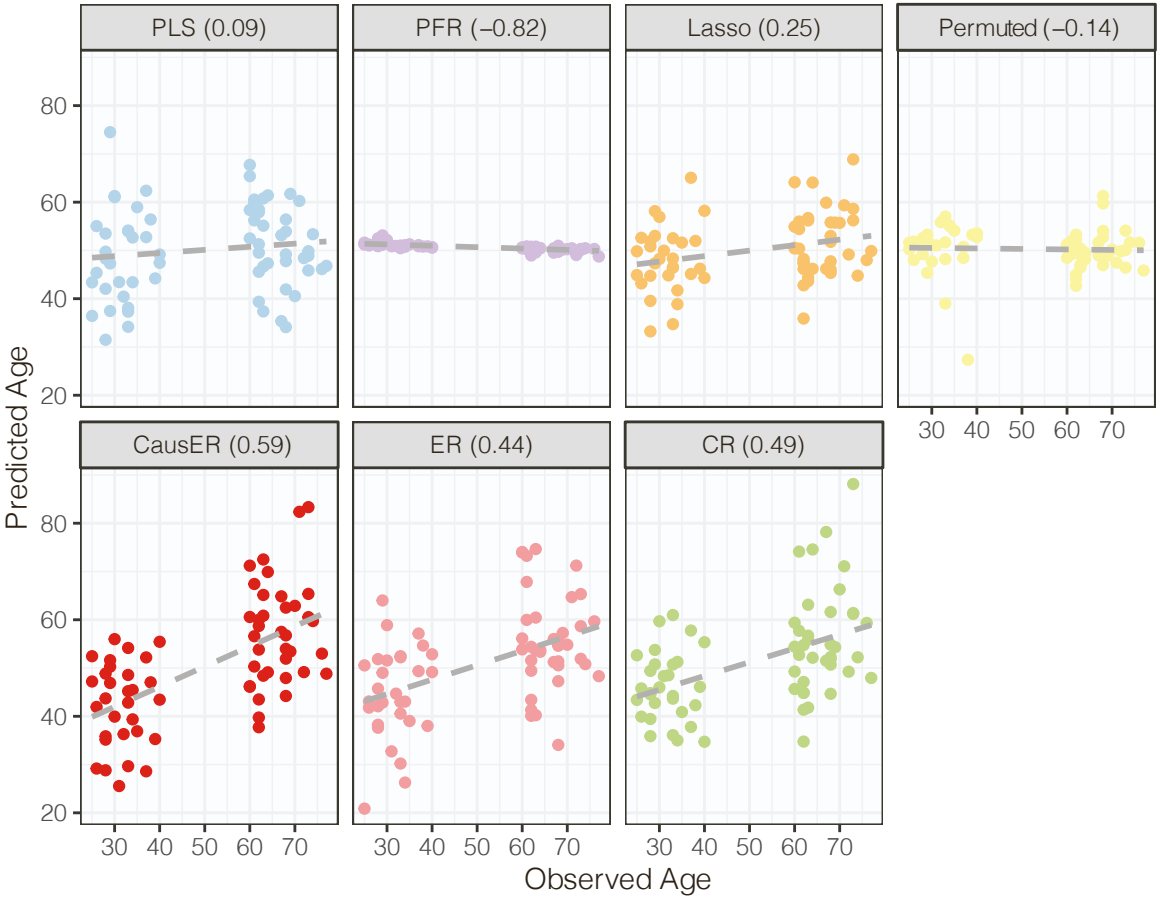

PLS (0.09)
PFR (−0.82)
Lasso (0.25)
Permuted (−0.14)
CausER (0.59)
ER (0.44)
CR (0.49)
Predicted Age
Observed Age
20
40
60
80
30
40
50
60
70

Fig. S1 – **Spearman correlations of the different methods at predicting age as a continuous variable, as measured in a LOOCV cross-validation framework**.

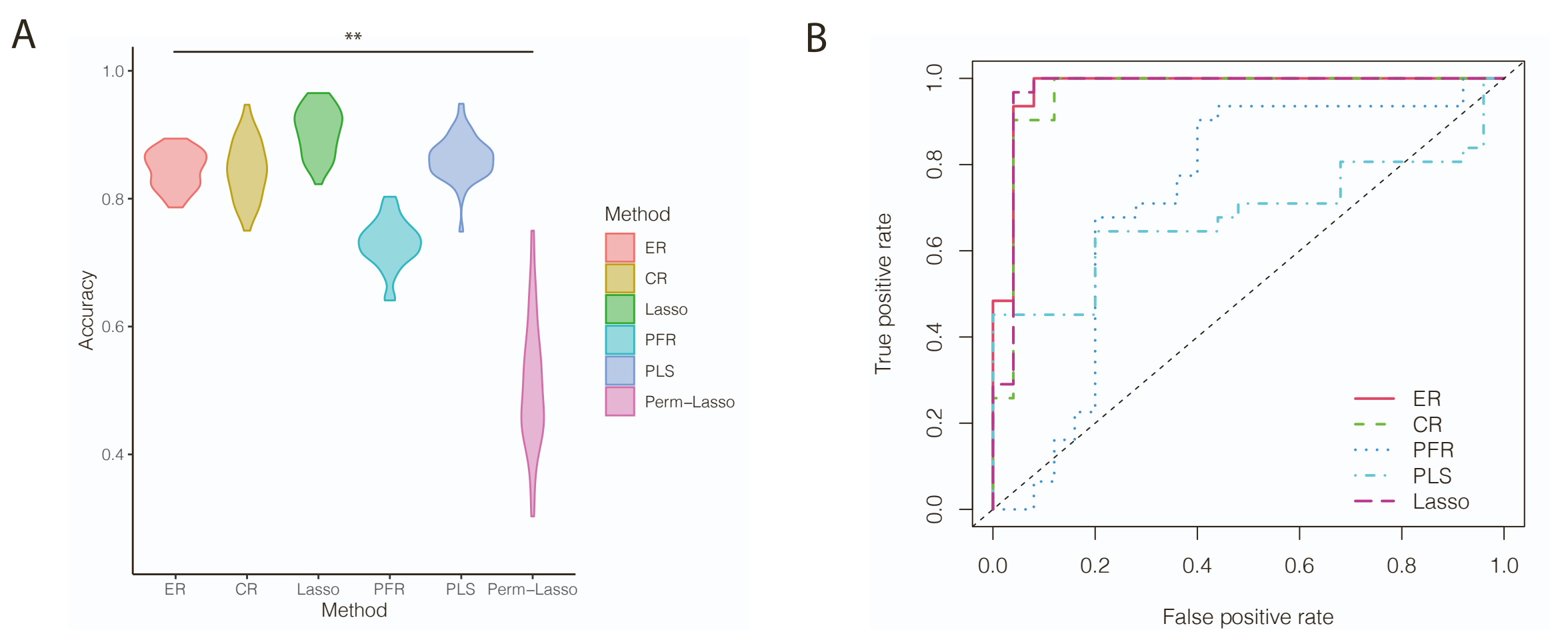

A
**
1.0
0.8
0.6
0.4
Accuracy
ER
CR
Lasso
PFR
PLS
Perm–Lasso
Method
Method
ER
CR
Lasso
PFR
PLS
Perm–Lasso
B
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
ER
CR
PFR
PLS
Lasso

Fig. S2 – Performance of ER in discriminating between term and pre-term births using immune profiles at week 1 after birth

a) Classification accuracy of the different methods at discriminating between term and pre-term births using immune profiles at 1 week after birth, measured in a replicated k-fold cross validation framework
b) ROC curves for the different methods at discriminating between term and pre-term births as measured in a LOOCV framework.

Supplementary Note S1

Detailed descriptions of the theoretical underpinnings, associated proofs of ER, CR and CausER. The file also describes details of the applications of ER, CR, CausER, LASSO, PLS and PFR to the different datasets of interest.

Supplementary Note 1

1 Latent factor model formulation

We assume that the data $(\mathbf{X}, \mathbf{Y}) \in (\mathbb{R}^{n\times p}, \mathbb{R}^n)$ are i.i.d. realizations of the random vector $(X, Y) \in (\mathbb{R}^p, \mathbb{R})$ which follows the factor regression model

$$Y = \beta^\top Z + \varepsilon \tag{1.1}$$
$$X = AZ + W. \tag{1.2}$$

The latent, unobserved, random vector $Z \in \mathbb{R}^K$, with $K < p$, is independent of the mean-zero random errors $\varepsilon \in \mathbb{R}$ and $W \in \mathbb{R}^p$. Independence between ε and W is assumed as well. The $p \times K$ matrix A and the K-dimensional vector β are deterministic quantities. We let $\mathbb{E}[\varepsilon^2] = \sigma^2$ and assume both $\mathbb{E}[ZZ^\top] = \Sigma_Z$ and A have rank K.

Since one of our main interests is to cluster the feature X based on its association with the latent factor Z, the membership matrix A needs to be identifiable, up to a $K \times K$ signed permutation matrix. For this reason, we resort to the following conditions on A, Σ_Z and $\Gamma := \mathbb{E}[WW^\top]$ (1).

Assumption 1.

(A0) $\|A_{j\cdot}\|_1 \le 1$ *for all* $j \in [p] := \{1, 2, \ldots, p\}$.

(A1) For every $k \in [K]$*, there exists at least two* $j \neq \ell \in [p]$*, such that* $|A_{j\cdot}| = |A_{\ell\cdot}| = e_k$.

(A2) There exists a constant $\nu > 0$ *such that*

$$\min_{1\le a<b\le K} \left([\Sigma_Z]_{aa} \wedge [\Sigma_Z]_{bb} - |[\Sigma_Z]_{ab}| \right) > \nu.$$

(A3) The covariance $\Gamma := \mathbb{E}[WW^\top]$ *is diagonal with bounded diagonal entries.*

Model (1.1) – (1.2) together with Assumption 1 is termed as the Essential Regression (ER). Within the ER framework, (1, Theorem 2) and (2, Proposition 1) establish that the model is identifiable. In particular, both A and β are identifiable. For the reader's convenience, we restate the results in Appendix A. The uniqueness of A is used to form unique clusters of X based on their associations with Z via

$$G_k := \{j \in [p] : A_{jk} \neq 0\}, \quad \forall k \in [K].$$

On the other hand, the coefficient β is used to select significant Z for predicting Y

To establish the identifiability results, the key condition is *(A1)* which assumes the existence of at least two features X that are *solely* associated with each latent factor Z. These features are termed as *non-mixed variables* and are collected in the set $I \subseteq [p] := \{1, 2, \ldots, p\}$. We further let $\mathcal{I} = \{I_1, \ldots, I_K\}$ denote its partition. Mathematically, we have

$$I_k := \{i \in [p] : |A_{ik}| = 1, A_{ik'} = 0, \text{ for all } k' \neq k\}, \quad \text{for } 1 \le k \le K.$$

The existence of non-mixed variables also renders the latent factors Z interpretable as the meaning of each Z_k can be read off from the corresponding non-mixed features $\{X_i\}_{i\in I_k}$.

Under the Essential Regression, our goals are two-fold:

(a) estimate β and select predictive latent factors Z;

(b) predicting Y.

In the next section, we will describe our approaches for each goal separately.

2 Methodology under Essential Regression

2.1 Estimation of β and selection of predictive Z

2.1.1 Estimation of β

Let $\widehat{\Sigma} := n^{-1}\mathbf{X}^\top\mathbf{X}$ denote the empirical sample covariance matrix of X. Our procedure for estimating β is the following.

(1) Obtain estimates $\widehat{K}$, $\{\widehat{I}_1, \cdots, \widehat{I}_{\widehat{K}}\}$, $\widehat{A}_{\widehat{I}\cdot}$ and $\widehat{\Sigma}_Z$ from $\widehat{\Sigma}$ with tuning parameter δ by using Algorithm 1 in (1). For the reader's convenience, the procedure is stated in Appendix B.

(2) Estimate $\Gamma_{\cdot I}$ by $\widehat{\Gamma}_{\cdot\widehat{I}}$ with

$$\widehat{\Gamma}_{ii} = \widehat{\Sigma}_{ii} - \widehat{A}_{i\cdot}^\top \widehat{\Sigma}_Z \widehat{A}_{i\cdot}, \quad \forall\, i \in \widehat{I}, \qquad \widehat{\Gamma}_{ji} = 0, \quad \forall\, j \neq i. \tag{2.1}$$

(3) Compute

$$\widehat{\Theta} = \left(\widehat{\Sigma}_{\cdot\widehat{I}} - \widehat{\Gamma}_{\cdot\widehat{I}}\right) \widehat{A}_{\widehat{I}\cdot} \left(\widehat{A}_{\widehat{I}\cdot}^\top \widehat{A}_{\widehat{I}\cdot}\right)^{-1}. \tag{2.2}$$

If $\widehat{\Theta}^\top\widehat{\Theta}$ is non-singular, estimate β by

$$\widehat{\beta} = \left(\widehat{\Theta}^\top\widehat{\Theta}\right)^{-1} \widehat{\Theta}^\top \frac{1}{n}\mathbf{X}^\top\mathbf{Y}. \tag{2.3}$$

Otherwise, compute

$$\widehat{h} = \frac{1}{n}\left(\widehat{A}_{\widehat{I}}^\top \widehat{A}_{\widehat{I}}\right)^{-1} \widehat{A}_{\widehat{I}}^\top \mathbf{X}_{\widehat{I}}^\top \mathbf{Y} \tag{2.4}$$

and estimate β by

$$\widehat{\beta}_d = \ \arg\min_{\beta\in\mathbb{R}^K} \left\{\|\beta\|_1 : \|\widehat{\Sigma}_Z\beta - \widehat{h}\|_\infty \leq \mu_1 + \mu_2\|\beta\|_1\right\} \tag{2.5}$$

for some parameters μ_1 and μ_2.

Algorithm 1 in step (1) requires to choose the tuning parameter δ. Since theoretical order of δ is $\sqrt{\log(p\vee n)/n}$ under the sub-Gaussian assumption of Z, ε and W, we set $\delta = c\sqrt{\log(p\vee n)/n}$ with the leading constant c chosen via the criterion in Section 5.1.1 of (1).

For $\widehat{\beta}_d$, the procedure requires additional tuning parmeters: μ_1 and μ_2. They are all of theoretical order $\sqrt{\log(p\vee n)/n}$. Our extensive simulation suggests to choose $\mu_1 = \mu_2 = 0.5\sqrt{\log(p\vee n)/n}$. Alternatively, they can be chosen via cross-validation by minimizing the loss

$$L(\beta) := \beta^T\widehat{\Sigma}_Z\beta - 2\beta^\top\widehat{h}.$$

2.1.2 Selection of predictive Z

When our estimator of β is $\widehat{\beta}$, (2, Theorem 4 and Proposition 5) provides the asymptotic distribution of $\widehat{\beta}_k$ for all $1 \leq k \leq K$ with consistent estimates of the asymptotic variances. We thus can construct confidence intervals (CIs) for each β_k, $1 \leq k \leq K$ and the obtained CIs could be used to select the predictive latent factors Z.

When our estimator of β is the Dantzig-type estimator $\widehat{\beta}_d$, as mentioned in (2, Remark 3 of version 1), $\widehat{\beta}_d$ adapts to the unknown sparsity of β. We propose to directly use the support of $\widehat{\beta}_d$ to select predictive Z.

2.2 Prediction

We have two procedures for predicting Y, described separately in the following two sections.

2.2.1 Essential regression predictor

For predicting Y, we adopt the procedure in (3, Section 4.2). Specifically, let $\widehat{\Theta}$ be constructed from (2.2) and compute

$$\widehat{\theta}_{ER} := (\widehat{\Theta}^\top \mathbf{X}^\top \mathbf{X} \widehat{\Theta})^- \widehat{\Theta}^\top \mathbf{X}^\top \mathbf{Y}$$

where M^- denotes the Moore-Penrose inverse of any matrix M. For any new data point (X^*, Y^*), we predict Y^* by

$$\widehat{Y}^*_{ER} = \widehat{\theta}_{ER}{}^\top X^*. \tag{2.6}$$

This predictor is termed as ER in our result.

2.2.2 Composite regression predictor

Within the Essential Regression framework, although the significant latent factors Z could be selected, the ER predictor in (2.6) still uses all the features X to predict Y. We thus propose a new predictor, called Composite Regression (CR), which uses only the features X that are related with the selected significant Z.

Specifically, CR has two steps. In the first step we select the significant factors Z and let $L \subseteq [K]$ be the index set of the selected Z. In the second step, we first find the subset of features X that are related with Z_L, that is, the set

$$\bar{S} := \left\{ j \in [p] : \|\widehat{A}_{jL}\|_2 \neq 0 \right\}$$

where $\widehat{A}$ is estimated from (B.4) in Appendix B, the procedure proposed in (1). We then regress $\mathbf{Y}$ onto $\mathbf{X}_{\bar{S}}$ via the Lasso approach to obtain the estimated linear coefficient vector

$$\widehat{\theta}_{CR} := \arg\min_{\theta} \|\mathbf{Y} - \mathbf{X}_{\bar{S}}\theta\|_2^2 + \lambda\|\theta\|_1.$$

The estimate $\widehat{\theta}_{CR}$ could be used to select predictive features associated with those significant factors Z. Furthermore, we propose

$$\widehat{Y}^*_{CR} := \widehat{\theta}^\top_{CR} X^* \tag{2.7}$$

to predict Y^*.

It is worth mentioning that the difference between CR and Lasso is that CR regresses $\mathbf{Y}$ onto $\mathbf{X}_{\bar{S}}$ based on the selected significant Z_L whereas Lasso regresses $\mathbf{Y}$ onto *all* the features $\mathbf{X}$. Hence the selected features X from CR are associated with the predictive factors, a desirable property that Lasso does not enjoy.

2.3 Prediction with Essential Regression on synthetic data

In this section[1] we generate synthetic data to compare the prediction performance of ER relative to PFR, PLS and the Lasso.

[1] This section is modified based on Section 5.2 in (4)

2.3.1 Data generating mechanism

We start with the description of our data generating mechanism. We first describe how we generate A, Σ_Z, Γ, and β. Recall that A can be partitioned into A_I and A_J.

To generate A_I, we set $|I_k| = m$ for each $k \in [K]$ and choose $A_I = \boldsymbol{I}_K \otimes \mathbf{1}_m$, where $\otimes$ denotes the kronecker product. Each row $A_{j\cdot}$ of A_J is generated by first randomly selecting its support with cardinality s_j drawn from $\{2, 3 \ldots, \lfloor K/2 \rfloor\}$ and then by sampling its non-zero entries from $N_{s_j}(0, D)$. The matrix D satisfies $\mathrm{diag}(D) = (1/s_j, \ldots, 1/s_j)$ and $D_{ab} = \zeta^{|i-j|}/s_j$ for any $a \neq b$ with given parameter $\zeta \in [0, 1]$. In the end, we rescale A_J such that the ℓ_1 norm of each row is no greater than 1.

To generate Σ_Z, we set $\mathrm{diag}(\Sigma_Z)$ to a K-length sequence from 2.5 to 3 with equal increments. The off-diagonal elements of Σ_Z are then chosen as $[\Sigma_Z]_{ij} = (-1)^{(i+j)}([\Sigma_Z]_{ii} \wedge [\Sigma_Z]_{jj})(0.3)^{|i-j|}$ for any $i \neq j \in [K]$. Finally, Γ is chosen by randomly sampling its diagonal elements from Unif(3, 5) and the entries of β are sampled independently from Unif(0, 1).

We generate the $n \times K$ matrix Z and the $n \times p$ noise matrix W whose rows are i.i.d. from $N_K(0, \Sigma_Z)$ and $N_p(0, \Gamma)$, respectively. Then we set $X = ZA^T + W$ and $Y = Z\beta + \varepsilon$ where the n components of ε are i.i.d. $N(0, 1)$. For each setting, we repeat generating (X, Y) 50 times and record the corresponding results.

Below we investigate how the prediction errors of ER, PFR, PLS and Lasso change as we vary p, K and the signal-to-noise ratio (SNR) one at a time. The performance metric is based on the new data prediction risk. To calculate it, we independently generate a new dataset (X_{new}, Y_{new}) containing n i.i.d. samples drawn according to our data generating mechanism. The prediction risk of the predictor $\widehat{Y}_{new}$ is calculated as $\|\widehat{Y}_{new} - Z_{new}\beta\|^2/n$.

2.3.2 Varying p, K and SNR one at a time

To vary p and K one at a time, we first set $n = 300$, $K = 10$, $m = 5$ and choose p from $\{200, 400, 600, 800, 1000\}$, then fix $n = 300$, $p = 600$, $m = 5$ and vary K in $\{10, 20, 30, 40, 50\}$. Both settings use $\zeta = 0.5$. We plot the prediction risks of the four predictors listed above.

To vary the signal-to-noise ratio $\xi = \lambda_K(A\Sigma_Z A^T)/\lambda_1(\Gamma)$, we fix Σ_Z and Γ, and generate A_J by choosing $\zeta \in \{0.1, 0.3, 0.5, 0.7, 0.9, 0.95, 0.99\}$. We set $n = 300$, $p = 400$, $K = 10$ and $m = 3$. For each ζ, we calculate the SNR and plot the prediction risks of each predictor.

Summary: Overall, the prediction error for all four methods deteriorates as K increases or the SNR decreases. This indicates that prediction becomes more difficult for large K and small SNR. On the other hand, ER, PFR and Lasso perform better as p increases. This contradicts the classical understanding that having more features increases the degrees of freedom of the model, hence inducing larger variance. By contrast, in our setting, increasing the number of features provides information that can be used to predict Z more accurately. This phenomenon has been observed in the classical factor (regression) model, see, for instance, (5–9).

Among the four candidates, ER has the smallest prediction error in all settings and PLS has the worst performance in most of the settings. Furthermore, PFR fails to detect K and tends to select $\widehat{K} < K$ in the second and third scenarios. It is clear that using $\widehat{K} < K$ leads to a large loss in prediction accuracy. This also indicates that, for principal component regression approaches, detecting K requires larger SNR than making consistent prediction with true K given. In the first plot, we are in a moderate SNR regime and PFR has comparable performance to ER. In the second plot, as K increases, the advantage of ER becomes considerable, which supports the fact that PFR only has guarantees for fixed K. Finally, in the third plot, the performance of PFR is more sensitive to the SNR comparing to the other three methods.

2.4 Non-linearity in Essential Regression

Our current model postulates two levels of linearity: the one between the observable feature vector X and the un-observed latent factor Z as $X = AZ+W$, and the one between the response Y and Z as $Y = Z^\top\beta + \varepsilon$. The linearity between observable feature X and latent factor Z is commonly assumed in the literature of factor models. Regarding the linearity between Y and Z, one could hope that this holds approximately as the dimension of Z grows (recall that neither Z nor its dimension K is known). This is part of the reasons why linear model is popular and appealing for modelling high-dimensional data (the feature dimension p is large). Moreover, both our algorithm and theoretical results are still valid for a moderate and large K, even when K grows with the sample size n. Therefore, the linearity between Y and Z is less restrictive when the number of latent factors is allowed to be large.

On the other hand, admittedly, linearity between Y and Z becomes restrictive in several cases, such as when Y is a categorical or binary response. This work is the first step towards making inference on β when Y and Z are linearly related. However, It is promising to extend the current framework to a more general setting where Y and Z follow generalized linear models. For example, our ongoing project studies the estimation and inference of the coefficient β when Y and Z are linked via logistic regression as

$$\text{logit}\left[\mathbb{P}(Y = 1|Z)\right] = Z^\top\beta.$$

When Y and Z are non-linearly related, a different procedure is needed in general.

2.5 Key methodological contributions of Essential Regression

We consider latent factor regression models, specified as

$$Y = Z^\top\beta + \varepsilon, \qquad X = AZ + W. \tag{2.8}$$

Note that Z is the un-observed latent factor, X is the observable feature and Y is the response. Our novel contribution is to develop valid inferential tool of the coefficient β, which can be used to select predictive *un-observed latent factors* (at the Z level). This is in stark contrast with the usual feature selection problem (for instance, studied by the Lasso) which aims to select predictive *features* (at the X level). Inference at the factor level is often more appealing in the multi-omics study than inference at the feature level to better understand the mechanism.

On the other hand, traditional factor (regression) models cannot be used to select predictive latent factors due to the lack of identifiability of β. In fact, another our main contribution is to establish practical conditions under which the coefficient β can be uniquely identified from the observable quantities of Y and X. Some existing conditions in the factor literature under which β becomes identifiable assume that the factors Z are uncorrelated, which is unrealistic in many applications.

To the best of our knowledge, the proposed method is the first paper that studies the identifiability, estimation and inference of β under model (2.8) where the latent factors Z are allowed to be fully correlated. Finally, it is also worth mentioning that, due to the existence of non-mixed variables, the latent factors become interpretable and their meaning can be read off from the corresponding non-mixed features. This interpretability makes the inference at the Z level meaningful. By contrast, the factors under existing conditions in factor model literature are not interpretable, leading their selection to be meaningless.

2.6 Comparison with topic models

Topic models are commonly used for learning thematic low-dimensional representations of text data. The postulated model is a particular instance of non-negative matrix factorizations. Concretely, in topic model, the observed data matrix $\mathbf{X}$ satisfies

$$\mathbf{X} = \mathbf{Z}A^\top + \mathbf{E}$$

where $\mathbf{X} \in \mathbb{R}^{p\times n}$, $A \in \mathbb{R}^{p\times K}$ and $\mathbf{Z} \in \mathbb{R}^{K\times n}$ all have non-negative entries and unit column sums. The non-negative constraint of both the data ($\mathbf{X}$) and the parameters of interest (A and $\mathbf{Z}$) renders the methods for topic models inapplicable to a more general setting where all of $\mathbf{X}$, $\mathbf{Z}$ and A are allowed to have negative entries. Moreover, the computationally efficient methods for topic models, such as (10–12), all focus on the identifiability and estimation of A without providing any inferential results.

3 Methodology under CausER

CausER is a novel method that combines the discovery of significant latent factors from Essential Regression with the causal inference implemented in CausalMGM, a method for learning causal graphs over mixed continuous and categorical data (13). This addresses two of the largest challenges of performing causal inference on biological datasets: high dimensionality and multicollinearity among variables. Highly collinear features are grouped into individual latent factors by the LOVE algorithm B, and the significant latent factor selection performed by ER 2.1.2 reduces the dimensionality of the dataset without discarding any latent factors causally linked to the response variable that we wish to predict or make inferences about.

3.1 Methodology under CausalMGM

3.1.1 Mixed Graphical Models

A Mixed Graphical Model (MGM) is an undirected graphical model capable of representing the joint distribution over datasets containing both continuous and categorical variables (14). The model is given by:

$$p(x, y; \theta) \propto \exp\left(\sum_{s=1}^{p}\sum_{t=1}^{p} -\frac{1}{2}\beta_{st}x_s x_t + \sum_{s=1}^{p}\alpha_s x_s + \sum_{s=1}^{p}\sum_{j=1}^{q}\rho_{sj}(y_j)x_s + \sum_{j=1}^{q}\sum_{r=1}^{q}\phi_{rj}(y_r, y_j)\right), \tag{3.1}$$

where θ represents the full set of parameters, x_s is the sth of p continuous variables, and y_j is the jth of q categorical variables. The parameter β_{st} represents the edge potential between continuous variables x_s and x_t, α_s represents the node potential of continuous variable s, ρ_{sj} represents the edge potential between continuous variable x_s and categorical variable y_j, and Φ_{rj} represents the edge potential between categorical variables y_j and y_r. A non-zero edge potential indicates the presence of an edge in the graph, and thus a pairwise conditional dependence relationship between those two variables. This model has the favorable property that the conditional probabilities of each variable can be represented with a Gaussian linear regression and multinomial logistic regression for continuous and categorical variables respectively. While learning this model over high-dimensional datasets directly is intractable due to the computation of the partition function, the above property enables us to learn the graphical model by minimizing the negative log-psuedolikelihood, given by 3.2, where Θ refers to the full set of model parameters.

$$\tilde{\ell}(\Theta \mid x, y) = -\sum_{s=1}^{p} \log p(x_s \mid x_{\setminus s}, y; \Theta) - \sum_{j=1}^{q} \log p(y_j \mid x, y_{\setminus j}; \Theta) \tag{3.2}$$

In order to ensure sparsity in the final model, we use proximal gradient descent to fit a penalized version of the negative log-pseudolikelihood from (15), given in 3.3. We identify the optimal values for regularization parameters λ_{CC}, λ_{CD}, and λ_{DD} using Stable Edge-specific Penalty Selection (StEPS), a method based on model stability, as defined in (15).

$$\underset{\Theta}{\text{minimize}}\, \ell_\lambda(\Theta) = \tilde{\ell}(\Theta) + \lambda_{CC} \sum_{s=1}^{p} \sum_{t=1}^{s-1} |\beta_{st}| + \lambda_{CD} \sum_{s=1}^{p} \sum_{j=1}^{q} ||\rho_{sj}||_2 + \lambda_{DD} \sum_{j=1}^{q} \sum_{r=1}^{j-1} ||\phi rj||_F \tag{3.3}$$

The undirected MGM does not represent a causal graphical model. However, it does identify pairwise conditional dependence relationships, and the resulting adjacencies are a *superset* of the adjacencies in the underlying causal graph. In the asymptotic sample limit, MGM learns the 'moralized graph', which consists of all edges in the causal DAG as well as additional edges between all spouses (nodes that share the same children) in the causal DAG. This means that the adjacencies learned from MGM can be used as the initial set of adjacencies for constraint-based causal inference methods, rather than a fully connected graph. This reduces the search space of the constraint-based causal inference algorithms, resulting in faster causal inference and fewer errors.

3.1.2 Fast Causal Inference

The FCI algorithm (16) learns a causal partial ancestral graph (PAG) from data that may include latent variables. Similar to the PC algorithm, the FCI algorithm performs conditional independence tests to find the skeleton of the final causal graph and orient the colliders (17). To accommodate the possibility of latent confounders in the causally insufficient case, the FCI algorithm must perform additional conditional independence tests that condition on some non-adjacent variables. This is because in the causally sufficient case with no latent confounders, if two variables are conditionally independent they are independent given some subset of their neighbors in the causal graph. However, this is no longer true in the causally insufficient case. The sets of variables that may cause two adjacent variables to be conditionally independent in the presence of latent confounders is characterized as the Possible D-Sep set. During the Possible D-Sep phase of the FCI algorithm, edges are pruned from the graph if the two variables are conditionally independent given one of the sets of variables in the Possible D-Sep set. Finally, all edges are reoriented to have circle endpoints, and the FCI orientation rules are applied as given by (18). In this paper, we use a version of the FCI orientation rules known as FCI-Max, where the initial collider orientation stage is done based on the separating set with the largest p value, as in (13). This heuristic, based on the observation the p values increase monotonically as the conditional dependence decreases (19), has been shown to considerably improve orientation accuracy over the initial implementation of FCI (13).

Constraint-based causal inference algorithms such as FCI require reliable conditional independence tests, which determine if some variables X and Y are independent given a conditioning set S, to learn the causal graphical model. Under the null hypothesis that X and Y are independent given S, we expect to see that $P(X \mid Y, S)$ is equal to the null model, $P(X \mid S)$. To accommodate mixed datasets as in (20), we compute these conditional probabilities with either linear regression, in the case that X and Y are continuous, or multinomial logistic regression, in the case that X, Y, or both are categorical. In the case that the test is performed by a linear

regression, we perform a t test on the coefficient of Y. If the p value of the test is less than the significance threshold α, then we reject the null hypothesis of conditional independence between X and Y given S. Alternatively, in the case that the test is performed by a multinomial logistic regression, we perform a likelihood ratio test (LRT) to determine whether $P(X \mid Y, S)$ is equal to the null model, $P(X \mid S)$. Again, if the p value of the test is less than the significance threshold α, then we reject the null hypothesis of conditional independence between X and Y given S. Note that the conditioning set S can contain both continuous and categorical variables, where categorical variables are transformed into an array of binary indicator variables.

The graphical causal models learned by FCI are PAGs, which are a representation of ancestry in causal graphs that is valid in the presence of latent confounders. Edges in this type of graph have three different types of endpoints: (o, $>$, -), and each represents an ancestral relationship between nodes in the graph. For example, $X \to Y$ indicates that X is an ancestor of Y, while $X \leftrightarrow Y$ indicates that a latent confounder causes both X and Y. The circular endpoint indicates uncertainty about the true causal endpoint. Thus, X o$\to Y$ could be either $X \to Y$ or $X \leftrightarrow Y$ in the true graph, meaning that the only certain knowledge according to the PAG is that Y is not an ancestor of X. An extensive theoretical grounding of these concepts can be found in (17).

3.1.3 Markov blanket

To build a predictor of a given response variable that is informed by the causal structure of variables in the dataset, a small subset of variables known as the Markov blanket are used. This set of variables are the variables that, when conditioned on, make the response variable independent of every other variable in the dataset according to the structure of the causal graph. In the case of a DAG, this set simply contains the parents, children, and spouses (other parents of the children) of the response variable (21). In the case of a PAG, the presence of latent confounders complicates the issue. In addition to the parents, children, and spouses of the response variable, we must include variables linked to the response variable or its children by a latent confounder ($X \leftrightarrow Y$) or a possible latent confounder (*e.g.*, X o$\to Y$), as well as the parents of those variables. If there is a chain of edges denoting the presence of latent confounders or possible latent confounders (*e.g.*, $W \leftrightarrow X \leftrightarrow Y \leftrightarrow Z$) that is linked to the response variable or its children, then the parents of each node in the chain is included in the Markov blanket (22). However, in practice this can lead to large Markov blankets and a drop in predictive performance, so we set the limit on the number of consecutive variables linked by latent confounders to include to 2. In the previous chain for example, if W was the response variable or one of its children, then X, Y, and their parents would be included in the Markov blanket.

By this definition, the Markov blanket of a response variable is the minimum set of features that contains all of the information available in the dataset for predicting the response variable. This can result in highly interpretable predictive models with few features and predictive performance that typically matches or exceeds those of models built with the full dataset or variables selected by Lasso. This has previously been successfully applied in biomedical applications, such as predicting whether lung nodules detected in low-dose CT scans are cancerous (23).

3.2 Constructing a causal model on significant latent factors

To identify the latent factors to use in the construction of a causal model, we perform Essential Regression on the variable of interest. In the case that $\widehat{K}$ is small compared to n, we learn a causal model over all latent factors identified by LOVE, the response variable, and any categorical variables that we wish to include in the model using CausalMGM. In the case that $\widehat{K}$ is large compared to n, we use the Dantzig estimator $\widehat{\beta}_d$ in 2.5 to identify latent factors that are significantly associated with the response variable. We then construct a causal model with the

latent factors that have non-zero coefficients in $\widehat{\beta}_d$, the response variable, and any categorical variables we wish to include in the model using CausalMGM.

When constructing the causal model, we first learned an undirected graphical model with MGM 3.1.1 (or GLASSO (24) if fully continuous). The optimal regularization parameters were selected based on graph stability using StEPS (15) (or StARS (25) if fully continuous). The resulting undirected graph was then used as an initial graph for performing causal inference with the FCI algorithm 3.1.2. This yields a causal PAG over the significant latent factors, the response variable, and any additional categorical variables, which can be used for the construction of predictive models or inference about causal mechanisms.

3.3 Stability-based α threshold selection

When building predictive models, the main structural feature of interest in the causal graph is the Markov blanket of the response variable. To select an optimal α threshold value for the conditional independence tests performed by the FCI algorithm, we took a stability-based approach based on StARS. While StARS was originally used for selecting the regularization parameter to be used in GLASSO, we use the same method of subsampling, learning the graph structure, and calculating the instability across subsamples to select an optimal α threshold for learning the Markov blanket of the response variable. However, instead of calculating the instability of an edge in the graph, we calculate the instability of a variable's inclusion in the Markov blanket.

We define $\widehat{\theta}_j(\alpha)$ as the frequency of a variable j's membership in the Markov blanket. Using this, we can calculate instability of a single variable j's membership in the Markov blanket as

$$\widehat{\xi}_j(\alpha) = 2\widehat{\theta}_j(\alpha)\left(1 - \widehat{\theta}_j(\alpha)\right). \tag{3.4}$$

This definition of the instability is twice the variance of the Bernoulli indicator for variable j's membership in the Markov blanket. Additionally, it can be interpreted as the probability of the Markov blankets learned on any two subsamples disagreeing about variable j's membership in the Markov blanket. This arises from the probabilities of the two possibilities for disagreement between subsamples: the probability that the first subsample includes variable j in the Markov blanket and the second subsample excludes variable j can be given as $\widehat{\theta}_j(\alpha)\left(1 - \widehat{\theta}_j(\alpha)\right)$, while the reverse can be given as $\left(1 - \widehat{\theta}_j(\alpha)\right)\widehat{\theta}_j(\alpha)$. When summed, this gives our definition of instability for a single variable j's membership in the Markov blanket. With this, we define the instability of the Markov blanket as a whole, $\widehat{D}(\alpha)$, as

$$\widehat{D}(\alpha) = \frac{\sum_{j \in MB} \widehat{\xi}_j(\alpha)}{m}, \tag{3.5}$$

where MB is the set of variables that shows up in the Markov blanket of the response variable in at least one subsample and at least one value of α, and m is the size of set MB. Very high values of α will lead to very dense but also very stable graphs, which is undesirable. To avoid this, we monotonize the instability of the Markov blanket as done in StARS, giving

$$\overline{D}(\alpha) = \sup_{0 \le t \le \alpha} \widehat{D}(t). \tag{3.6}$$

As it it more difficult to test the instability of many α parameters with CausalMGM and CausER than with the regularization λ in GLASSO, we modify the selection of the optimal α threshold

$\widehat{\alpha}$. While StARS selects the smallest value of λ with an instability less than some threshold γ, we select the value of α with an instability closest to some threshold γ, given by

$$\widehat{\alpha} = \inf_{\alpha} |\overline{D}(\alpha) - \gamma|. \tag{3.7}$$

This method for selecting $\widehat{\alpha}$ requires the selection of a threshold γ. This may make the method seem redundant, as we require a new hyperparameter in order to select $\widehat{\alpha}$. However, this threshold γ has a clear interpretation in the context of the learned graph; it represents the average probability of graphs learned on two subsamples disagreeing on the membership of a variable in the Markov blanket. In contrast, the stability of the graph can vary considerably at the same values of α in different datasets. Thus, by setting the threshold $\gamma = 0.05$, we are selecting the value of α that results in Markov blanket selections where the average probability that a variable is present in one selection but not in another is closest to 0.05.

3.4 Causally informed prediction of the response variable

Once the optimal threshold value $\widehat{\alpha}$ is selected with the above method, we build a final causal model using the full dataset and the conditional independence test threshold $\widehat{\alpha}$. We then identify the Markov blanket of the final causal model to be used as predictors in the construction of regression models for the response variable. If the response variable is continuous, we use linear regression, and if the response variable is categorical we use multinomial logistic regression. Predictive performance is estimated using leave-one-out cross-validation, where we train the model on all but one sample and then predict the value of the held out sample. This procedure is repeated for each sample in the dataset, and the performance metric is calculated using the predictions of the held out values.

4 Specifications of the data analysis

We provide detailed specifications of all the data analysis carried out in this paper.

4.1 Imputation of missing values

Among the datasets that we studied, there are different levels of missingness. To impute the missing values, we use the averaged value of the 5 nearest neighbors in Euclidean distance.

4.2 Implementation of different methods

Throughout our analysis, we consider the following competitive methods:

1. CausER: CausER in Section 3.
2. ER: Essential Regression in (2.6) of Section 2.1.
3. CR: Composite Regression in (2.7) of Section 2.1. The tuning parameter of the Lasso step uses the k-fold cross validation. We use $k = 10$ if there are more than 3 observations per fold, otherwise set $k = 5$.
4. Lasso: The Lasso (26) from `glmnet` implemented in `R` with the tuning parameter λ_{lasso} selected from the k-fold cross validation with k chosen according to the same rule for CR. When Lasso selects no feature, we randomly select 5 features and use an ordinary least squares estimator based on these 5 features.

5. PFR: principal factor regression which regresses $\mathbf{Y}$ on the first K principal components of $\mathbf{X}$ where K is selected based on the criterion proposed in (27, 28). Specifically, we estimate K by

$$\widehat{K} = \arg \max_{k \in \{1,2,\ldots,\bar{K}\}} \frac{\widehat{\lambda}_k}{\widehat{\lambda}_{k+1}} \tag{4.1}$$

where $\widehat{\lambda}_1, \widehat{\lambda}_2, \cdots$ are the non-decreasing eigenvalues of $\mathbf{X}^\top \mathbf{X}/n$ and $\bar{K}$ is some prespecified value, for instance, the largest integer that is no greater than $\min(n, p) - 1$.

6. PLS: partial least squares regression from `plsr` implemented in `R` with the number of components selected by the default function `selectNcomp`.

4.3 Cross-validated assessment of predictive performance

Two cross-validation techniques were used to assess the predictive performance of the different methods: (1) replicated 10-fold cross-validation, and (2) leave-one-out cross-validation.

1. **Replicated 10-fold cross-validation:** For assessing the accuracy of the classifiers in the RTS,S vaccine-induced transcriptomic profiles dataset and the Term / pre-term infants stereotypic immune convergence dataset, 50 replicates of nested 10-fold cross-validation was performed. On each fold, in each replicate, we independently ran each of the methods in 4.2 and assessed the predictive accuracy. For ER, the latent factors were learned on each fold and each replicate, and the regression and final latent factor selection were repeated. For CausER, a causal model was learned over the latent factors selected as significant by ER for each fold and replicate. The average cross-validation accuracy across the 10 folds was calculated for each of the 50 replicates.

2. **Leave-one-out cross-validation:** For all three datasets, we performed leave-one-out cross-validation to assess the accuracy of each method. In leave-one-out cross-validation, each sample in the dataset is held out as the predictive models are trained on the remaining $n - 1$ samples, and then the held out sample is predicted with the trained models. The assessment of model performance is done with the set of predictions of the left out values. These predictions were used to calculate Receiver Operating Characteristic (ROC) curves, correlations between predictions and true values, and classification accuracies.

4.4 Multi-omic responses to the Zostavax vaccine dataset

To construct the dataset of multi-omic responses to the Zostavax vaccine, we included the following multi-scale measurements of immune state: IgG titers, blood transcriptional modules, metabolic clusters, $CD4^+$ T cell populations, T_{FH} cell populations, flow cytometry cell populations, cytokine profiles, and IFNγ T cells. We used subject age as the response variable for the n = 72 subjects. We exclude the features that have missing values for more than a half of subjects. We also exclude 5 subjects that have no observed features. The remaining data sets are merged via the unique id's of subjects. The final data set contains $p = 1721$ features of $n = 67$ subjects.

We applied ER with $\delta = 0.38$ and obtained $\widehat{K} = 57$ clusters. As $\widehat{K}$ is relative large comparing to n, we used the Dantzig estimator $\widehat{\beta}_d$ of β in (2.5) and the non-zero support $\widehat{\beta}_d$ selects 18 significant factors for predicting the response.

For this dataset, which is fully continuous, the initial undirected skeleton was learned with Graphical LASSO (GLASSO) (24) implemented in the `huge` package in `R` (29). The optimal regularization parameter $\lambda = 0.32$ was selected by StARS (25). The causal model over

the latent factors was built using FCI-Max, as implemented in the `rCausalMGM` package in `R` (https://github.com/tyler-lovelace1/rCausalMGM). The optimal conditional independence test significance threshold for learning the causal graph, $\alpha = 0.1$, was selected as described in 3.3.

4.5 Term / pre-term infants stereotypic immune convergence dataset

For pre-processing the data, we first combine the data sets of Cell population frequencies and Short final ComBat by removing the irrelevant features such as "gender", "mode of delivery", "family" etc. Then we further exclude the control samples with row indices from 326 to 337. We further pulled out the subdata with "Relation" equal to "child" and the final data set we use has $n = 183$ samples and $p = 282$ features with 56 samples from week 1 and 46 samples from week 12. The response is binary, either "Control" (representing term) or "Premature" (representing pre-term). We use the 5-NN to impute the missing values.

We perform the classification of pre-term / term by using the features collected in week 12. We applied ER with $\delta = 0.15$ and obtained 14 clusters. Our estimator $\widehat{\beta}$ is constructed via (2.3) and the 95% confidence intervals select two significant factors Z_5 and Z_7 for predicting term and pre-term.

For features in week 1, we used $\delta = 0.17$ with 14 clusters and the significant factors include Z_3, Z_4, Z_{10}, Z_{11} and Z_{14}.

Only week 12 data was analyzed with CausER. For this dataset, only two significant latent factors, Z_5 and Z_7, were identified, making causal orientations unidentifiable in most cases (the only exception being the graph $Z_5 \circ\!\!\rightarrow Y \leftarrow\!\!\circ Z_7$). Additionally, there are too few features for stability-based selection of α using only the significant latent factors. However, a causal model was constructed over all latent factors using FCI-Max, as implemented in the `rCausalMGM` package in `R` (https://github.com/tyler-lovelace1/rCausalMGM). The optimal conditional independence test significance threshold for learning the causal graph, $\alpha = 0.2$, was selected as described in 3.3.

4.6 RTS,S vaccine-induced transcriptomic profiles dataset

By concatenating the two gene-expression data sets, we end up with $n = 116$ samples with $p = 22277$ probes. We filtered out the probes that could map to multiple genes, and then the technical replicates were averaged with the `limma` package in `R` (30), giving the expression of $p = 12424$ genes.

The responses $Y \in \mathbb{R}^n$ are categorical representing three time points. We applied ER to the data set with selected $\delta = 0.04$ and obtained $\widehat{K} = 1674$ clusters. The estimator of β is the Dantzig estimator in (2.5) which has 86 non-zero elements. This implies there are 86 significant factors Z for predicting the response.

For this dataset, which is mixed, the initial undirected skeleton was learned with MGM, described in 3.1.1, implemented in the `rCausalMGM` package in `R`. The optimal regularization parameters $\lambda_{CC} = 0.27$, $\lambda_{CD} = 0.27$ (λ_{DD} was irrelevant because there is only one categorical variable) were selected by StEPS (15). The causal model over the latent factors was built using FCI-Max, as implemented in the `rCausalMGM` package in `R` (https://github.com/tyler-lovelace1/rCausalMGM). The optimal conditional independence test significance threshold for learning the causal graph, $\alpha = 0.2$, was selected as described in 3.3.

4.7 Antibody glycosylation in active / latent tuberculosis dataset

This dataset consists of measurements of glycosylation in bulk non-antigen-specific IgG, bulk Fc domain, bulk Fab domain, and purified protein derivative (PPD)- and Ag85A-specific IgG from patients with latent ($n = 10$) and active ($n = 20$) tuberculosis. The dataset also contains PPD-specific isotype, PPD-specific antibody dependent phagocytosis, cellular cytotoxicity, and natural killer cell activation. In total, this dataset has $n = 30$ samples and $p = 181$ features and a binary response variable representing either latent or active tuberculosis.

We applied ER with $\delta = 0.35$ to perform the classification of active/latent tuberculosis. We obtained 8 clusters with only one significant factor, Z_7, for predicting latent vs. active tuberculosis.

We also analyzed this dataset with CausER. For this dataset, we did not learn an initial undirected graph due to the low dimensionality of the latent factors ($K = 8$) and low sample size. Instead, we directly constructed a causal model over all latent factors and the response variable Y using FCI-Max, as implemented in the `rCausalMGM` package in `R` (https://github.com/tyler-lovelace1/rCausalMGM). The optimal conditional independence test significance threshold for learning the causal graph, $\alpha = 0.1$, was selected as described in 3.3.

5 Computational complexity of Essential Regression and CausER

5.1 Computational complexity for Essential Regression

The primary computational cost of ER is the feature clustering step of the algorithm. This step requires the computation of the full covariance matrix for the features in the dataset, and thus scales according to $\mathcal{O}(np^2)$. The regression step of ER either uses least squares or a K-dimensional linear program, and thus is fast for small K.

5.2 Computational complexity for learning MGM

The causal discovery algorithm used here, CausalMGM, enables causal discovery on large datasets by first learning an undirected graphical model, MGM, through proximal gradient descent. Let the sample size be n and the number of latent factors identified by ER be K. Then the computational complexity of the MGM algorithm scales according to $\mathcal{O}(nK^2)$.

5.3 Computational complexity for FCI-Max

The causal discovery algorithm used here, FCI-Max, is constraint-based, and thus needs to perform large numbers of conditional independence tests. Thus, the runtime of FCI-Max and other constraint-based algorithms are dependent on the runtime of the conditional independence test, and the number of conditional independence tests that needs to be performed. The number of conditional independence tests that needs to be performed, in turn, depends on the structure of the causal graph. Thus, we can only give an upper bound for the worst-case running time, which will be presented here; in practice, the runtime is typically much lower than this upper bound.

Let the sample size be n, the number of latent factors be K, and the maximal degree of the causal graph be d. Then, in the worst case scenario, we get this upper bound on the number of conditional independence tests:

$$2\binom{K}{2}\sum_{i=0}^{d}\binom{K-1}{i} \leq \frac{K^{d+2}}{d!} \tag{5.1}$$

Dataset	n	p	K	ER (s)	MGM (s)	FCI-Max (s)
Zostavax shingles vaccine	67	1721	57	4.71	0.398	0.025
Term/Pre-term	46	274	14	0.6	N/A	0.021
RTS,S malaria vaccine	116	12424	590	75.6	0.722	0.734
Active/latent tuberculosis	30	181	8	0.433	N/A	0.003

Table 1: Wall clock runtimes (in seconds) of all three components of Essential Regression and CausER in the four datasets analyzed here.

While this runtime is a high order polynomial, meeting the assumptions on the structure of the graph needed to achieve the worst case is highly unlikely. Nonetheless, when combined with the runtimes of our regression-based conditional independence tests, which are $\mathcal{O}(nd^2 + d^3)$, we get the following upper bound on computational complexity:

$$\mathcal{O}\left(\frac{(nd^2 + d^3)K^{d+2}}{d!}\right) \tag{5.2}$$

When we learn an undirected graphical model as a skeleton, as is done with MGM here, we can significantly reduce the worst case number of conditional independence tests and therefore this upper bound. Let $|E_0|$ be the number of edges in the initial graph, and d_0 be the maximal degree of of the initial graph, where $d \leq d_0$. Under the worst case scenario, we get this upper bound on the number of conditional independence tests:

$$2|E_0| \sum_{i=0}^{d} \binom{d_0}{i} \leq \frac{2|E_0|(d_0 + 1)^d}{d!} \tag{5.3}$$

When combined with the runtimes of our regression-based conditional independence tests, we get the following upper bound on computational complexity:

$$\mathcal{O}\left(\frac{(nd^2 + d^3)|E_0|(d_0 + 1)^d}{d!}\right) \tag{5.4}$$

As mentioned above, the assumptions on the true causal graph and set of conditional independence results required to result in the worst-case running time is highly unlikely. To meet these assumptions in the case where no initial undirected graph is provided, every node in the causal graph must be of the maximal degree d, and every pair of non-adjacent variables must only be conditionally independent with conditioning sets of size d. When an initial undirected graph is provided, in addition to the prior two conditions, every node in the undirected graph must have the maximal degree of the initial graph, d_0. Thus, runtimes in practice are typically well below the upper bound given here, as can be seen from the empirical results of the runtimes on the datasets studied here.

A Identifiability results of A and β

We re-state the identifiability results of A and β from (1) and (2).

Theorem 1 (Theorems 1 & 2 (1)). *Under model $X = AZ + E$ with Assumption 1, the set of pure variable I, its partition $\mathcal{I} = \{I_1, \ldots, I_K\}$[2] and the number of factors are identifiable from $\Sigma = Cov(X)$.*

Moreover, the matrix A is identifiable up to a $K \times K$ signed permutation matrix.

Proposition 2 (Proposition 1 (2)). *Under model (1.1) – (1.2) with Assumption 1, the coefficient vector β is identifiable up to a signed permutation matrix.*

B The LOVE algorithm

We first give the specifics of estimating I and K developed by (1).

Algorithm 1 Estimate the partition of the pure variables $\mathcal{I}$ by $\widehat{\mathcal{I}}$

procedure PureVar($\widehat{\Sigma}$, δ)
  $\widehat{\mathcal{I}} \leftarrow \varnothing$.
  for all $i \in [p]$ **do**
    $\widehat{I}^{(i)} \leftarrow \{l \in [p] \setminus \{i\} : \max_{j \in [p]\setminus\{i\}} |\widehat{\Sigma}_{ij}| \le |\widehat{\Sigma}_{il}| + 2\delta\}$
    $Pure(i) \leftarrow True$.
    for all $j \in \widehat{I}^{(i)}$ **do**
      if $\big||\widehat{\Sigma}_{ij}| - \max_{k \in [p]\setminus\{j\}} |\widehat{\Sigma}_{jk}|\big| > 2\delta$ **then**
        $Pure(i) \leftarrow False$,
        break
    if $Pure(i)$ **then**
      $\widehat{I}^{(i)} \leftarrow \widehat{I}^{(i)} \cup \{i\}$
      $\widehat{\mathcal{I}} \leftarrow$ Merge($\widehat{I}^{(i)}$, $\widehat{\mathcal{I}}$)
  return $\widehat{\mathcal{I}}$ and $\widehat{K}$ as the number of sets in $\widehat{\mathcal{I}}$

function Merge($\widehat{I}^{(i)}$, $\widehat{\mathcal{I}}$)
  for all $G \in \widehat{\mathcal{I}}$ **do** ▷ $\widehat{\mathcal{I}}$ is a collection of sets
    if $G \cap \widehat{I}^{(i)} \neq \varnothing$ **then**
      $G \leftarrow G \cap \widehat{I}^{(i)}$ ▷ Replace $G \in \widehat{\mathcal{I}}$ by $G \cap \widehat{I}^{(i)}$
      return $\widehat{\mathcal{I}}$
  $\widehat{I}^{(i)} \in \widehat{\mathcal{I}}$ ▷ add $\widehat{I}^{(i)}$ in $\widehat{\mathcal{I}}$
  return $\widehat{\mathcal{I}}$

Next, for each $a \in [\widehat{K}]$ and $b \in [\widehat{K}] \setminus \{a\}$, we compute

$$\big[\widehat{\Sigma}_Z\big]_{aa} = \frac{1}{|\widehat{I}_a|(|\widehat{I}_a| - 1)} \sum_{i,j \in \widehat{I}_a, i \neq j} |\widehat{\Sigma}_{ij}|, \qquad \big[\widehat{\Sigma}_Z\big]_{ab} = \frac{1}{|\widehat{I}_a||\widehat{I}_b|} \sum_{i \in \widehat{I}_a, j \in \widehat{I}_b} \widehat{A}_{ia}\widehat{A}_{ib}\widehat{\Sigma}_{ij}, \tag{B.1}$$

to form the estimator $\widehat{\Sigma}_Z$ of Σ_Z. Furthermore, we restate the estimation of $A_{I\cdot}$ in (1). For each $k \in [\widehat{K}]$ and the estimated pure variable set $\widehat{I}_k$,

$$\text{Pick an element } i \in \widehat{I}_k \text{ at random, and set } \widehat{A}_{i\cdot} = e_k; \tag{B.2}$$

$$\text{For the remaining } j \in \widehat{I}_k \setminus \{i\}, \text{ set } \widehat{A}_{j\cdot} = \operatorname{sign}(\widehat{\Sigma}_{ij}) \cdot e_k. \tag{B.3}$$

[2] $\mathcal{I}$ is identifiable up to a group permutation.

For the estimation of $A_{J\cdot}$, we use the Dantzig-type estimator $\widehat{A}_D$ proposed in (1) given by

$$\widehat{A}_{j\cdot} = \arg\min_{\beta^j} \left\{ \|\beta^j\|_1 : \ \left\| \widehat{\Sigma}_Z \beta^j - (\widehat{A}_{\widehat{I}\cdot}^\top \widehat{A}_{\widehat{I}\cdot})^{-1} \widehat{A}_{\widehat{I}\cdot}^\top \widehat{\Sigma}_{\widehat{I}j} \right\|_\infty \le c\sqrt{\log(p \vee n)/n} \right\} \tag{B.4}$$

for any $j \in \widehat{J}$, with some constant $c > 0$. The estimator $\widehat{A}$ enjoys the optimal convergence rate of $\max_{j\in[p]} \|\widehat{A}_{j\cdot} - A_{j\cdot}\|_q$ for any $1 \le q \le \infty$ (1, Theorem 5).

References

1. X. Bing, F. Bunea, Y. Ning, M. Wegkamp, *Ann. Statist.* **48**, 2055–2081 (Aug. 2020).
2. X. Bing, F. Bunea, M. Wegkamp, *In arXiv:1905.12696* (2019).
3. X. Bing, F. Bunea, S. Strimas-Mackey, M. Wegkamp, *Prediction in latent factor regression: Adaptive PCR and beyond*, 2020, arXiv: 2007.10050 (stat.ML).
4. X. Bing, F. Bunea, M. Wegkamp, S. Strimas-Mackey, *arXiv: 1905.12696* (2019).
5. J. H. Stock, M. W. Watson, *Journal of the American Statistical Association* **97**, 1167–1179, ISSN: 01621459 (2002).
6. J. Bai, *Econometrica* **71**, 135–171 (2003).
7. J. Bai, S. Ng, *Journal of Econometrics* **146**, Honoring the research contributions of Charles R. Nelson, 304–317 (2008).
8. J. Bai, S. Ng, *Econometrica* **74**, 1133–1150 (2006).
9. J. Fan, Y. Liao, M. Mincheva, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**, 603–680 (2013).
10. S. Arora *et al.*, presented at the Proceedings of the 30th International Conference on Machine Learning, ed. by S. Dasgupta, D. McAllester, vol. 28, pp. 280–288, (https://proceedings.mlr.press/v28/arora13.html).
11. X. Bing, F. Bunea, M. Wegkamp, *Bernoulli* **26**, 1765–1796 (2020).
12. X. Bing, F. Bunea, M. Wegkamp, *Journal of machine learning research* **21** (2020).
13. V. K. Raghu *et al.*, *International journal of data science and analytics* **6**, 33–45 (2018).
14. J. D. Lee, T. J. Hastie, *Journal of Computational and Graphical Statistics* **24**, 230–253 (2015).
15. A. J. Sedgewick, I. Shi, R. M. Donovan, P. V. Benos, *BMC bioinformatics* **17**, 307–318 (2016).
16. P. L. Spirtes, C. Meek, T. S. Richardson, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 499–506 (1995).
17. P. Spirtes, C. N. Glymour, R. Scheines, D. Heckerman, *Causation, prediction, and search* (MIT press, 2000).
18. J. Zhang, *Artificial Intelligence* **172**, 1873–1896 (2008).
19. J. Ramsey, *arXiv preprint arXiv:1610.00378* (2016).
20. A. J. Sedgewick *et al.*, *Bioinformatics* **35**, 1204–1212 (2019).
21. C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, X. D. Koutsoukos, *Journal of Machine Learning Research* **11** (2010).
22. J. Zhang, *Journal of Machine Learning Research* **9**, 1437–1474 (2008).
23. V. K. Raghu *et al.*, *Thorax* **74**, 643–649 (2019).
24. J. Friedman, T. Hastie, R. Tibshirani, *Biostatistics* **9**, 432–441 (2008).
25. H. Liu, K. Roeder, L. Wasserman, *Advances in neural information processing systems* **24**, 1432 (2010).
26. R. Tibshirani, *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 267–288, ISSN: 00359246 (1996).

27. S. C. Ahn, A. R. Horenstein, *Econometrica* **81**, 1203–1227, eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA8968 (2013).

28. C. Lam, Q. Yao, *Ann. Statist.* **40**, 694–726 (Apr. 2012).

29. T. Zhao, H. Liu, K. Roeder, J. Lafferty, L. Wasserman, en, *J. Mach. Learn. Res.* **13**, 1059–1062 (Apr. 2012).

30. M. E. Ritchie *et al.*, *Nucleic acids research* **43**, e47–e47 (2015).