

Generative prediction of causal gene sets responsible for complex traits

Benjamin Kuznets-Speck^{a,b}, Buduka K. Ogonor^{a,b}, Thomas P. Wytock^{a,b}, and Adilson E. Motter^{a,b,c,d,e,f,1}

Affiliations are included on p. 9.

PNAS

Edited by H. Mark Johnston, University of Colorado-Anschutz Medical Campus, Aurora, CO; received July 26, 2024; accepted May 9, 2025

The relationship between genotype and phenotype remains an outstanding question for organism-level traits because these traits are generally complex. The challenge arises from complex traits being determined by a combination of multiple genes (or loci), which leads to an explosion of possible genotype-phenotype mappings. The primary techniques to resolve these mappings are genome/transcriptome-wide association studies, which are limited by their lack of causal inference and statistical power. Here, we develop an approach that combines transcriptional data endowed with causal information and a generative machine learning model designed to strengthen statistical power. Our implementation of the approach-dubbed transcriptome-wide conditional variational autoencoder (TWAVE)-includes a variational autoencoder trained on human transcriptional data, which is incorporated into an optimization framework. Given a trait phenotype, TWAVE generates expression profiles, which we dimensionally reduce by identifying independently varying generalized pathways (eigengenes). We then conduct constrained optimization to find causal gene sets that are the gene perturbations whose measured transcriptomic responses best explain trait phenotype differences. By considering several complex traits, we show that the approach identifies causal genes that cannot be detected by the primary existing techniques. Moreover, the approach identifies complex diseases caused by distinct sets of genes, meaning that the disease is polygenic and exhibits distinct subtypes driven by different genotype-phenotype mappings. We suggest that the approach will enable the design of tailored experiments to identify multigenic targets to address complex diseases.

gene regulatory networks | biological networks | nonlinear dynamics | complex systems | generative deep learning

Complex traits are polygenic, orchestrated by networks of interacting genes that work together to produce phenotypic variation (1-3). An outstanding question in the study of such traits is the identification of the specific combinations of gene variants that give rise to the different phenotypic expressions (4-10). Association studies have been performed to search for genetic loci significant to a complex trait phenotype by conducting hypothesis tests on *individual* genetic loci, from which *independent* mutations/genes are statistically associated with the phenotype in question (11-15). We innovate on these techniques by developing a framework to *jointly* predict sets of genes while accounting for collective behavior not captured by statistical tests on individual genes.

Association studies, such as genome/transcriptome-wide association studies (GWAS/ TWAS), have been broadly adopted in over 5,700 studies and 3,300 traits as of 2021 (12). A common critique of these studies is that they have low statistical power due to the combinatorial explosion in the number of gene sets that must be tested (11, 13, 16). Post-GWAS/TWAS analyses such as *fine-mapping* attempt to address this limitation by considering the correlation structure of the genetic data (17-21). However, they rely on an initial association study to select what variants to fine-map, potentially leaving behind genes that would be significant collectively but have low independent effect size. In the framework presented here, we develop and apply an approach that considers all genes simultaneously, regardless of their individual effect size. The framework combines generative machine learning, dimensionality reduction, and constrained optimization (Fig. 1).

A key aspect of our approach is the use of increasingly available trait-labeled transcriptomic data from bulk and single-cell RNA-Seq experiments, which contends with the biological networks that influence complex traits (22, 23). To better extract patterns from our transcriptional data, we develop the transcriptome-wide conditional variational autoencoder (TWAVE), a generative deep learning model that generates

Significance

Researchers have long sought to bridge the gap between phenotypes and the genotypes that cause them. This gap remains open because current methods focus on associating phenotypes to a combinatorially explosive number of genotypic possibilities, resulting in a loss of statistical power. We overcome this limitation by employing transcriptomic data from complex, polygenic, human diseases combined with measured transcriptomic responses to gene perturbations in cell lines. The former data allow us to perform generative modeling and dimensional reduction to map transcriptome to phenotype, while the latter incorporate causal information regarding how gene regulation shapes phenotype. We predict sets of genes that explain the emergence of complex traits, which suggest possible multitarget disease treatments.

This article is a PNAS Direct Submission

Copyright © 2025 the Author(s). Published by PNAS. This open access article is distributed under Creative Commons Attribution License 4.0 (CC BY)

¹To whom correspondence may be addressed. Email: motter@northwestern.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas. 2415071122/-/DCSupplemental.

Published June 12, 2025.

Author contributions: B.K.-S., T.P.W., and A.E.M. designed research; B.K.-S. and B.K.O. performed research; B.K.-S. and T.P.W. analyzed data; and B.K.-S., B.K.O., T.P.W., and A.E.M. wrote the paper.

The authors declare no competing interest.



Fig. 1. Schematic of the proposed approach. Synthetic transcriptomic profiles are generated from a machine learning model that learns from RNA-Seq data on complex traits. The generated data are projected onto eigengenes, which are linear combinations of genes that vary independently, retain important gene correlations, and differentiate between complex trait phenotypes. From here, gene perturbations whose transcriptional responses bridge the gap between trait phenotypes are found by constrained optimization.

denoised transcriptional profiles for the relevant phenotypes. To understand how regulatory changes affect expression levels, we also incorporate complementary data on transcriptional response to gene perturbations (knockdowns and overexpressions) (24). These transcriptional data are dimensionally reduced while maintaining causal information, which we achieve using the concept of eigengenes, to facilitate the optimization over gene perturbations (25). Together, these data sources allow us to explore how the regulatory network drives phenotypic changes without prior knowledge of network structure.

We focus on the human disease traits in Table 1. Throughout, we take care to distinguish between *traits* (e.g., eye color) and their *phenotype* variants (e.g., blue, brown, green), and we consider traits that have a baseline and variant phenotype. Moreover, we interpret the states associated with each variant as defined by distinct attractors of the gene regulatory network (26–28). The problem of identifying the genes that cause a trait phenotype can thus be mapped to an optimization over combinations of transcriptional perturbations that steer transcriptomic states from baseline to variant attractors and vice-versa. The resulting framework reveals groups of gene perturbations that most influence phenotypic variation, pinpointing the molecular underpinnings that determine complex traits.

Results

Generating Complex Trait Transcriptomes. Identifying *sets* of differentially expressed genes is complicated by the fact that transcriptional measurements include a large number of genes and a comparatively small number of samples. It is precisely this feature that makes it challenging to distinguish real biological differences from random variance when using statistical tests that treat genes as independent variables. Our method recognizes that genes operate in concert rather than independently to orchestrate cell function, which transforms the problem into learning an effective representation of these relationships from data. Fig. 2 presents TWAVE, which solves this problem by looking at the data as a whole, using a neural network encoder to embed high-dimensional gene expression profiles onto a low-dimensional

latent space (Z), where data points can be classified and new representative points can be generated. Points in the latent space, including newly generated ones, are decoded back up to the full gene expression space as in Fig. 2A. The model (consisting of the encoder, decoder, and latent space classifier) is trained with a combination of three loss terms. The first term accounts for how accurately the autoencoder can reconstruct the original data. The second term is the Kullback–Leibler (KL) divergence loss, which regularizes Z so that each dimension contributes roughly equally to the overall variance in the latent space. The third term is a classification loss to mold the structure of the latent space so that different phenotypes of the trait (baseline and variant for example) can be distinguished by their transcriptional states, as detailed in *Materials and Methods*.

Fig. 2B shows that transcriptional measurement associated with the baseline and variant populations (blue and red clusters) segregate to coordinates along the first principal component, which is a natural outcome of the latent space learned by TWAVE. Here, we use the representative case of inflammatory bowel disease to illustrate the construction of TWAVE, but we observe similar performance among the other traits we considered in Table 1. The figure also shows that a linear interpolation between the two clusters in Z-space lies along the first principal component, which accounts for the largest fraction of variation in the data. Fig. 2C demonstrates a close agreement between the original gene expression distributions and the reconstructions from TWAVE. To test whether our autoencoder retains associations between genes and the complex trait of interest, we compare the differentially expressed genes identified by t tests on both the original and reconstructed expression profiles using the area under the receiver operating characteristic curve (AUROC). Fig. 2D shows this curve, which is constructed by 1) arranging both sets of genes from smallest to largest in terms of their *p*-values, 2) varying the threshold at which genes are statistically significant, and 3) counting the fraction of significant genes in the original data that are selected by TWAVE (true positive rate) as a function of the fraction of genes selected by TWAVE that are *not* significant in the original data (false positive rate). For all traits, we find that the AUROC approaches 1, indicating that the sets of differentially expressed genes identified by TWAVE and within the original data are nearly identical. Full technical details concerning the construction of TWAVE are provided in Materials and Methods.

For each of the complex trait datasets described in Table 1, we use TWAVE to estimate the distributions of the transcriptional data in the latent space arising from the baseline and variant phenotypes, while retaining the fundamental features that distinguish the two populations. We then draw points from these distributions in the latent space, decoding them as depicted in Fig. 2A. By choosing an equal number of each phenotype (baseline and variant), estimates of the distributions from the data are equally precise for each of these trait phenotypes. It is instructive to compare against the method of "extreme pseudosampling," in which the latent space of a variational autoencoder (VAE) is sampled randomly (29, 30). Our method is different from extreme pseudosampling in two crucial ways. First, TWAVE employs a conditional VAE (i.e., it includes a latent space explicitly trained to classify between baseline and variant). Second, TWAVE draws from a probability distribution in the latent space associated with the trait phenotype label instead of drawing randomly from any state in the latent space. Overall, TWAVE allows us to make the most from limited transcriptional data, drawing new representative samples in a way that would be unfeasible without generative modeling.

Complex trait	Tissue	Seq. type(s)	N _{baseline}	N _{variant}	GEO series
Allergic asthma	Peripheral blood mononuclear cells	Bulk	277	166	GSE96783
Inflammatory bowel disease	Gastrointestinal tissue	Bulk	461	2029	GSE193677
Food allergy	CD4+ T cells	Bulk	71	63	GSE114065
Cancer metastasis	Metastasis: Breast \rightarrow lung	Single cell	1170	1274	GSE202695
Macular degeneration	Macular retina and retinal pigment epithelium	Bulk & single cell	433	104	GSE135092
Type 1 diabetes	CD4+ T cells	Single cell	557	2502	GSE182870
Non-small cell lung cancer	Blood platelets	Bulk	376	400	GSE89843
Simple trait	Tissue	Seq. type	N _{baseline}	N _{variant}	GEO series
MODY3	Differentiated embryonic stem cells	Single cell	113	158	GSE129653
Complex trait	Tissues	Seq. type	N _{baseline}	N _{variant}	DepMap series
Pancancer metastasis	86 primary cancers, 34 lineages	Single cell	838	447	24Q2

The columns represent (from Left to Right) the traits considered, the tissues of origin, the type of sequencing, the number of samples of the baseline and variant phenotypes, and the GEO or DepMap series accession numbers.

Causal Dimensions of Complex Trait Variation. We select the *eigengenes* that are most determinative of the trait phenotypes (i.e., are causal). Conceptually, eigengenes are weighted combinations of genes that vary in concert within the eigengene, but any given eigengene can vary independently of the others. Mathematically, they are eigenvectors of the TWAVE-generated gene expression matrix *Y* and form an orthogonal basis e_i in gene expression space (25). This basis corresponds to the columns of the unitary matrix $U = [e_1, ..., e_i, ..., e_l]$ in the singular value decomposition

 $Y^T = U\Sigma V^T$, where Σ is a diagonal matrix of singular values in descending order and V^T contains the left eigenvectors of Y^T . For each dataset, we perform singular value decompositions of the $m \times n$ matrix Y, where m is the number of sample expression profiles in the TWAVE dataset, n is the number of genes in each sample, and $l = \min(m, n)$ is the rank of Y.

We proceed to determine which eigengenes are most likely to capture differences between the baseline and variant trait phenotype by adapting Bayesian fine-mapping (17) to eigengenes. This



Fig. 2. TWAVE construction and validation, presented for the inflammatory bowel disease trait. (A) TWAVE architecture, where gene expression profiles are projected onto a low-dimensional latent space (Z) and subsequently reconstructed with a neural network decoder. (B) First two principal components (PC) of the latent space Z, showing a clear separation of complex trait phenotypes (baseline and variant). A linear interpolation between the means of the two populations in the latent space (blue-to-red stars) falls along the first principal component. (C) Comparison between original (blue) and TWAVE-reconstructed (red) distributions of gene expression for four different genes, conveying strong agreement. (D) Receiver operating characteristic for significant gene associations in the erginal data.



Fig. 3. Dimension reduction by selecting the most causal eigengenes for the inflammatory bowel disease trait. (*A*) Posterior inclusion probability for eigengenes to be causal arranged in descending order of their singular values. (*B*) Classification accuracy from logistic regression on data projected onto an increasing subset of the top *d* eigengenes arranged in descending order of the sosterior probability of being causal (blue), in descending order of the singular values (gray), and randomly (red).

procedure seeks a small set of r eigengenes that can accurately distinguish between the phenotypes according to the posterior inclusion probability, which quantifies how well a proposed set of eigengenes explains the data (i.e., how causal the set is). The fine-mapping procedure starts by projecting the data onto the d = 200 eigengenes with the largest singular values $X = YU_{n \times d}$. This choice of d ensures that the set of eigengenes from which the causal set is selected accounts for the large majority of the variance, as shown in SI Appendix, Fig. S1. A logistic regression model over eigengenes is then fit using the expression profiles and associated trait phenotype labels, achieving high accuracy, F1-score, and recall, for all datasets analyzed. We also show that a maximum likelihood estimator for the posterior distribution of causal eigengenes can be formed from: 1) the odds ratios $\zeta = \log[\rho/(1-\rho)]$ from the logistic regression, where ρ is the probability of a data point belonging to the variant class; and 2) the projected expression matrix X. To reduce from d to reigengenes, we perform Markov chain Monte Carlo (MCMC) sampling to maximize the likelihood of causal eigengenes given the regression data. For each of these d eigengenes, we evaluate the posterior inclusion probability that each eigengene is causal $p(e_i \text{ causal} | \text{data} = \{X, \zeta\})$ and build sets from the top r = 50causal eigengenes (r = 10 for allergic asthma) with the largest posterior inclusion probability. We show $p(\mathbf{e}_i|X,\zeta)$ for the first d principal components of Y in Fig. 3A, ordered from noncausal (p = 0) to causal (p = 1).

As a validation test for the selected eigengenes, we perform the logistic regressions on X summarized in Fig. 3B. The regression accuracy for including data projected onto the first *i* eigengenes sorted in order of causality $p(e_i|X, \zeta)$ quickly climbs to above 0.9 within *r* top eigenvectors. This validates our choice of keeping *r* eigengenes in our dimensionality reduction. On the other hand, including eigengenes in principal component order (i.e., ordered by the fraction of variance that aligns along each eigenvector), yields significantly poorer classification results. Arranging the

eigengenes randomly can actually produce a better result than arranging them in order of the singular values when keeping less than 60 eigengenes, but principal component ordering outperforms the random one as the dimension of this reduced space is increased.

Complex Trait Transitions via Eigengene Perturbations.

Constrained optimization. To implicate genes responsible for transitioning between the baseline and variant phenotypes, we explore extensive data on transcriptional responses to gene perturbations. Specifically, we define a perturbation response matrix, *B*, whose rows represent eigengenes in the original dataset and whose columns are average transcriptional responses to a transcriptional perturbation (one response profile for each column). This matrix consists of 10% overexpressions and 90% knockdowns (most of the latter are implemented through RNA interference), as indicated in Dataset S1. With this matrix in hand, we investigate which combinations of perturbations can cause the baseline transcriptional profile to match the variant, and vice-versa. Formally, this question is answered by solving the following constrained optimization problem:

$$\boldsymbol{u}^* = \operatorname*{argmin}_{\boldsymbol{u}|0 \leq u_{\alpha} \leq 1} \{ D(\boldsymbol{u}) + \lambda \sum_{\alpha} u_{\alpha} \}, \qquad [1]$$

where

$$D(\boldsymbol{u}) = ||\boldsymbol{x}_{\text{variant}} - \boldsymbol{x}_{\text{baseline}} - B\boldsymbol{u}||, \qquad [2]$$

and $||\cdot||$ denotes the Euclidean distance. The choice of Euclidean distance reflects our assumptions 1) that there is a single phenotype for each transcriptional state and 2) that differences in the expression of each eigengene are equally likely to contribute to phenotypic differences. We discuss alternative choices of the distance metric in *SI Appendix*. Here, perturbations add with weight \boldsymbol{u} in causal eigengene space to transition from the state $\boldsymbol{x}_{\text{baseline}}$ to the state $\boldsymbol{x}_{\text{variant}}$ (Fig. 4*A*). Before considering transitions between individual states in the baseline and variant clusters, we consider transitions between *average* baseline and variant states.

Since we expect the baseline and variant phenotypes to be stable with respect to the fluctuations inherent to transcription, the closer we approach states known to belong to a given phenotype transcriptionally, the more likely that state is to exhibit that phenotype. We quantify this likelihood with the coefficient of determination

$$R^2 = 1 - D(u^*)/D(0),$$
 [3]

where $R^2 \leq 1$ and R^2 close to 1 indicates high efficacy. In particular, $R^2 > 0.5$ indicates that the distance between the two states has been at least halved. The relationship between distance and cell behavior becomes less precise in the full high-dimensional expression space because there are many points that are a given distance away from a target point. Thus, reducing the number of dimensions of our problem by working with a select set of eigengenes is crucial, which is implemented through our choice to express matrix *B* in the space of eigengenes.

We use the Python function minimize from SciPy (which implements the L-BFGS-B method) to solve the constrained optimization problem in Eq. 1 for all of our RNA-Seq datasets. The optimal u^* yields $R^2 \approx 1$ for all complex traits we consider, namely allergic asthma, inflammatory bowel disease, food allergy, cancer metastasis (where baseline and variant refer to primary and metastatic tumors), age-related macular degeneration, type 1 diabetes, and non-small cell lung cancer (Fig. 4*B*). This is further



Fig. 4. Attributing complex trait phenotypes to sets of genes. (*A*) Identification of gene sets that differ between baseline and variant trait phenotypes, where blue and red dots represent individual states. A state in the baseline cluster (blue background) transitions to a state in the variant cluster (red background) upon targeted transcriptional perturbations. (*B*) Coefficient of determination for controlled transitions between phenotypes of complex traits, where R^2 close to 1 indicates that the final transcriptional state approaches that of the target phenotype.

confirmed by examining the selected perturbations for individual trait phenotypes and noting that they generally point toward the phenotype expression in question.

Table 2 shows the genes with the top 12 perturbation weights in u^* for allergic asthma along with a brief annotation of their function. Many of these top-selected genes have been implicated in allergic asthma, lung and airway function, and inflammation and immunity. Mutations in BMPR2, for example, have been shown to cause asthma-like symptoms and pulmonary hypertension in response to mild antigens in the airway (34, 35). In addition, TCF7 promotes T-cell differentiation to Th2 or memory T cells (36), consistent with allergic asthma being the result of immune system dysregulation. Other identified genes, such as TARDBP, TENT4B, and HNRNPL, have not been previously implicated in allergic asthma. However, these genes are associated with RNA metabolism and modifications including alternative splicing and poly-A tail alteration, which in turn are related to immune response (46). In particular, TARDBP (TDP-43) has been shown to regulate alternative splicing and alternative polyadenylation in CD8+ T cells, and specific RNA splicing and polyadenylation events depend on the presence of TARDBP during CD8+ T-cell costimulation (31). TENT4B is involved in mRNA stabilization, influencing B-cell proliferation and the cellular response to viral infections (32), whereas HNRNPL participates in the regulation of inflammatory responses, particularly through its interaction with long noncoding RNAs and its role in regulating *TNF-\alpha* transcription (44, 45). We performed the optimization between the average baseline and variant states for all 6 other complex traits shown in Fig. 4B, and refer the reader to SI Appendix, Tables S1-S6 for the top selected genes for each trait.

Optimization for individual baseline-variant states. We perform optimizations across individual baseline and variant pairs to account for the fact that the measured transcriptional signatures of a given phenotype vary heterogeneously across cell samples and individuals (Fig. 4A). An analogous approach in TWAS would require subsampling an already small number of measurements, leading to a large uncertainty in the variance and an inability to detect baseline-variant differences (47). By recasting the hypothesis test as an optimization problem, we avoid this issue using information on how the regulatory network responds to perturbations. This reformulation allows us to investigate how the gene perturbations responsible for a trait may change across individual baseline-variant pairs. We break down our strategy into two steps: 1) find the optimal set of perturbations for a large number of baseline-variant state pairs and 2) compare the observed co-occurrence of perturbation pairs with a null model. The null model is designed to preserve both the frequency with which each perturbation is selected and the number of perturbations needed for each pair of states.

The transition between any x_i and x_j can be induced by applying the perturbation

$$\boldsymbol{u}_{ij} = \operatorname*{argmin}_{\boldsymbol{u} \mid 0 \le u_{\alpha} \le 1} \{ ||\boldsymbol{x}_i - \boldsymbol{x}_j - B\boldsymbol{u}|| + \lambda \sum_{\alpha} u_{\alpha} \}, \qquad [4]$$

where the variant state is \mathbf{x}_i and the baseline state is \mathbf{x}_j . We solve this optimization problem over a range of different λ values, taking the largest λ (the sparsest solution) such that $R^2 = 1 - D(\mathbf{u}_{ij})/D(\mathbf{0}) > 0.99$. This is repeated for N = 2,500 randomly selected pairs in the forward (baseline-to-variant) and reverse directions. We then construct a bipartite network represented by the (adjacency) matrix

$$A = [\dots, \boldsymbol{u}_{ij}, \dots]^T,$$
 [5]

where the columns of A are perturbed genes and the rows are different accepted *i*, *j* pairs. We take the dot-product between the columns μ and ν of A to get the frequency $f_{\mu\nu}$ at which the corresponding perturbed genes co-occur in the same u_{ij} :

$$f_{\mu\nu} = \frac{1}{N} \sum_{\alpha=1}^{N} A_{\alpha\mu} A_{\alpha\nu}.$$
 [6]

To identify statistically significant pairs μ , ν , we compare $f_{\mu\nu}$ to the frequency at which μ and ν occur together in a null model

Table 2.Allergic asthma transcriptional perturbations(GSE96783)

Gene	Annotation
TARDBP-	RNA metabolism and regulation (31)
TENT4B	Posttranscriptional modifications (32)
KRR1 [_]	Ribosome biogenesis (33)
BMPR2	Inflammatory signaling (34, 35)
TCF7 ⁺	Growth/migration in airway (36)
APOBEC3G ⁺	Innate immunity, antiviral (37)
INTS12 ⁻	Lung function via protein synthesis pathways (38)
NEAT1	Inflammation in asthma (39)
MTHFD1 ⁻	Folate and methionine metabolism (40, 41)
PRMT5 ⁻	Allergic airway inflammation (42)
FASTKD1 ⁻	Mitochondrial function, apoptosis (43)
HNRNPL	RNA splicing and expression regulation (44, 45)

of a random maximum entropy graph whose row and column sums are fixed to that of A (48, 49). The expected co-occurrence frequency in the null model is

$$\langle f_{\mu\nu} \rangle = \frac{1}{N} \sum_{\alpha=1}^{N} p_{\alpha\mu} p_{\alpha\nu}, \qquad [7]$$

where p_{ij} is the probability of an edge existing between nodes *i* and *j* of the maximum entropy graph (*Materials and Methods*). In addition to $\langle f_{\mu\nu} \rangle$, the SE $\sigma_{\mu\nu}$ can be approximated via error propagation using the fact that the probability of edge occurrences are Bernoulli random variables. From these ensemble statistics a *z*-value can be constructed as $z_{\mu\nu} = (f_{\mu\nu} - \langle f_{\mu\nu} \rangle)/\sigma_{\mu\nu}$.

Now that we have established a null model for the cooccurrence of perturbations in causing/reversing the variant behavior, we can examine the network formed by the statistically significant co-occurrences that deviate from the maximum entropy model. To determine where co-occurrences begin to deviate from the null model, we identify the set of significant pairs with high $z_{\mu\nu}$ above a threshold defined by inspection of the quantile–quantile plot (*SI Appendix*, Fig. S2).

Application to Allergic Asthma. In the case of allergic asthma, pairs with $z_{\mu\nu} > 20$ were kept for analysis. The corresponding network is depicted in Fig. 5A for the baseline-to-variant transition (i.e., the onset of asthma). Each node represents a perturbed gene within a significantly co-occurring pair and is color-coded according to the nature of the perturbation (knockdown or overexpression). Edges represent significant cooccurrences and are color-coded by the expression correlations across responses between the genes they connect. We find that genes with many connections in this network representation, such as ADAR, PAN3, and MAPK1 have been implicated in allergic asthma before (50-52). We also optimized over gene perturbations in the reverse direction, as shown in Fig. 5B. Among the genes featured in this network, we again find several linked to allergic inflammation: SUZ12 inhibition is associated with the reduction of allergic inflammation through is role in the protein complex PRC2, JAK2 inhibitors have been proposed to alleviate asthma because of JAK2's role in the JAK-STAT signaling pathway, and knockdown of MYC has been shown to repress ILC2 (type 2 innate lymphoid cell, a type of immune cell) activity, which in turn reduced airway inflammation and immune hyperresponsiveness (53-55).

Note that many of the genes in the forward and reverse cooccurrence graphs in Fig. 5 A and B are *distinct*. In a dynamical network, reversing a perturbation does not generally restore the state of the system, a phenomenon that is accounted for by the bounds placed on u in Eqs. 1 and 4. These bounds may prevent the same genes from being selected in the forward and reverse directions, which within our approximation reflects the fact that the responses to knockdowns and overexpressions are not exactly antialigned. In our case, this phenomenon gives rise to the observation that the genes causing a given trait phenotype are not necessarily the ones that mitigate it. Moreover, as shown in Fig. 5C, the number of perturbations required to make the transition in the forward and reverse directions are also different. Remarkably, it takes a combination of fewer singlegene perturbations to induce a transition in the reverse direction (i.e., from the variant to the baseline state) than in the forward one. This may be because it takes more perturbations to go to a particular variant state than to a generic baseline state ("all roads lead to Rome" but not necessarily the inverse). A second



Fig. 5. Genes perturbed in transitions between baseline and variant clusters for allergic asthma. (*A* and *B*) Gene perturbation co-occurrence networks for forward (*A*) and reverse (*B*) transitions. Edges appear between a pair of perturbed genes if they frequently co-occur in successful transitions (i.e., they both occur with high frequency in u_{ij} compared to a maximum entropy graph null model). The edges are colored by gene-gene correlation in the perturbation response dataset. The nodes are sized proportionally to the number of edges and color-coded according to whether the perturbation is a knockdown (blue) or overexpression (red). (*C*) Histograms of the number of gene perturbations required to induce the forward transition (red) and the reverse transition (blue).

possibility is that there is an overrepresentation of certain genes in the library of gene perturbations (*Materials and Methods*).

Importantly, both genes involved in the forward direction and genes involved in the reverse direction can be associated with the trait because they can promote or reverse a change from the baseline to variant phenotype. For example, the involvement of MYC in the reverse direction may be due to its known function in promoting plasticity between transcriptional states and amplifying gene expression overall when overexpressed (56, 57). Moreover, pluripotent stem cells with MYC knocked-down have been shown to decrease allergic reactions in mice by inhibiting T-helper cell immune reaction (58). Aberrant translation of the CEBPA gene, which is implicated in the forward direction, has also been associated with causing bronchial smooth muscle cells-a tissue that plays a key role in asthma-to proliferate faster (59). Notably, there are genes featured in both the forward and reverse co-occurrence networks that play roles in both promoting and attenuating allergic asthma. For example, FOXO1 overexpression in mice has been shown to promote allergic asthma through macrophage polarization, Th9 (T-helper 9 cell) differentiation, and regulation of IRF4 expression, though inhibition of *FOXO1* led to attenuation of immune response and asthmatic inflammation through regulation of IRF4 (60, 61). Likewise, the role of JUNB depends on the state of other transcription factors. Though JUNB significantly influences Th2 (T-helper 2 cell) differentiation and the production of Th2 cytokines, promoting allergic inflammation, it also plays a role in maintaining homeostasis. Specifically, knockdown of JUNB limits excessive inflammation by modulating regulatory T-cell differentiation (62).

We also find that, except for *BMPR2* and *TENT4B*, the genes appearing in the co-occurrence graphs from the optimization between *individual* baseline–variant states are distinct from those identified by the *average* variant-baseline optimization (Table 2). The apparent contrast between optimizing over average transcriptional states and individual pairs highlights the fact that there can be multiple paths (defined by different perturbation sets) through which the disease progresses and is mediated and that these paths are not necessarily the ones connecting the average transcriptional states. Consequently, the distinct co-occurring genes in Fig. 5 could potentially relate to different mechanisms playing a role in allergic asthma. We emphasize that such a cooccurrence structure *cannot* be inferred from studying population averages alone, as typically done in GWAS/TWAS (47).

It is instructive to consider the transcription factors that regulate the genes in our perturbation response library (upstream transcription factors). Not all upstream transcription factors have transcriptional responses measured in our library, so we use the Enrichr gene set enrichment database to find them. We focus on the case where the upstream factors simultaneously regulate both genes in a co-occurring pair so that a single transcription factor could explain their combined influence on the trait phenotype. For instance, GATA2, TET2, and TWIST1 are enriched for more than one gene co-occurrence pair and are known to influence allergic asthma (63-65). The most parsimonious explanation for the enrichment of these transcription factors is that they exert their influence on the phenotype (at least in part) through the genes that appear in our perturbation response dataset. The enrichment of these additional factors shows that we may be able to infer trait-associated genes outside our dataset.

Learning Across Different Contexts with TWAVE. Thus far, we have considered 7 complex traits, each based on data from a unique tissue type and previously examined by differential

expression. Next, we generalize to contexts where 1) the data (and trait phenotype) in question are associated with multiple disparate tissues and 2) the trait phenotype is caused by a mutation that affects the function of the protein instead of its transcriptional expression.

We consider a phenotype that manifests itself through multiple tissues by studying the trait of cancer metastasis in the cancer dependency map (DepMap) dataset (66). Because these *pancancer* data come from many different cell types, there are confounding variables that render a simple differential expression analysis unable to detect any differentially expressed genes common to the process of metastasis across all tissues. Indeed, we found no statistically significant associations by performing such an analysis. However, using TWAVE, we are able to disentangle the effects of the confounding variables associated with different disease contexts (e.g., cell type, tissue origin, tumor location, and systematic effects) to uncover common biological mechanisms driving cancer metastasis (SI Appendix, Fig. S3). In this case, we again find that our co-occurring perturbation networks contain many genes previously found to promote or mitigate cancer metastasis, including NF1 knockdown, SOX5 overexpression, CBFB overexpression, TOX4 knockdown, PROX1 overexpression, and EHF knockdown (67-72). An Enrichr search for the co-occurring genes reveals out-of-sample upstream transcription factors that are known to affect metastasis as well, such as STAT3 and CTCF (73, 74). Although both of these genes appear to be essential for growth (75, 76), which limits opportunities to perturb them in cell-line experiments, one can detect their influence on trait phenotype variation through the pairs of genes in the co-occurrence graph that they regulate.

To examine the scenario in which a causal mutation affects a gene's protein function but not its transcriptional expression, we consider maturity-onset diabetes of the young type 3 (MODY3). MODY3 is known to be a largely monogenic trait caused by mutations to the transcription factor HNF1A that impact beta cell function and diabetes in general. Since HNF1A is one of the genes perturbed in our perturbation response matrix B, MODY3 provides an excellent example where the solution is known. Documented mutations of HNF1A alter its protein function, which in turn alters the expression of *other* genes as opposed to its own. In fact, HNF1A overexpression appears in 30.2% of baselinevariant pair optimizations. This is compared to the top overall perturbation, NEAT1 knockdown, appearing 57.4% of the time. However, HNF1A, as opposed to NEAT1, also appears in the forward co-occurrence network (SI Appendix, Fig. S4). Of the top 13 gene perturbations, three of them-MED1 knockdown, HNF1A overexpression, and GATA2 overexpression-were also in the forward co-occurrence network and have been implicated in diabetic function.

This narrows down the large list of possible genes to a number that could be tested in low-throughput lab experiments. For instance, *MED1* knockout mice show a heightened sensitivity to insulin and an improved glucose tolerance (77). All of the other genes in the co-occurrence network exhibit transcriptional responses that are highly positively correlated with that of *HNF1A*, meaning that their corresponding column vectors in *B* all point in the same direction. This pattern is markedly different from those observed in the complex traits above, in which we also find transcriptional responses that are negatively correlated and uncorrelated. Among the perturbations correlated with *HNF1A* overexpression is *ALOX5* overexpression, which also impacts beta cell function in diabetes via increased insulin resistance (78, 79).

Finally, we directly compare the genes identified by our method, differential expression, and TWAS in the case of

inflammatory bowel disease in *SI Appendix*, Fig. S5. We find that the only 8% of the differentially expressed genes identified in the dataset (80) overlap with the TWAS genes, which is reflective of the challenges inherent to reconciling results produced by different approaches. Applying our method, we find that 36% of the genes participating in over 54% of the solutions to Eq. 4 overlap with TWAS. We emphasize that this improvement by our method occurs because it accounts for the downstream impacts of the gene perturbations through the *B* matrix, which naturally filters out spurious differentially expressed genes.

Discussion

The approach presented here leverages existing transcriptomic data to address the challenges that complex traits pose to traditional mutation-association screening methods (13). We implement this by identifying generalized cellular pathways (eigengenes) relevant to a complex trait and by calculating optimal sets of gene perturbations whose transcriptional responses change the combined state of these pathways from one phenotype to another. Our approach accounts for limited data, heterogeneity within phenotypes, confounding biological variation, and combinatorial explosion in gene sets in ways that traditional methods cannot (13). In particular, limited data are addressed by our development of the generative model TWAVE; inference of heterogeneous pathways is illustrated in the example of allergic asthma; common drivers of cancer progression across biological subtypes are found in the DepMap example; and finally, a combinatorial explosion is avoided by casting the identification of causal genes as an optimization problem. The approach can also implicate candidate genes through their known downstream effects on the gene regulatory network obtained from experiments.

An overarching goal of our approach is to narrow the scope of candidate gene combinations to a number amenable to targeted low-throughput experiments. As in previous successful applications of Boolean networks (81–86) and principal component-based techniques that uncover low-dimensional structure in gene regulatory networks (87), we aim to uncover causal influences. The main advantage of our approach is that it can generate predictions solely from publicly available data without explicit network reconstruction or specific knowledge concerning the gene functions and interactions, making it broadly applicable.

It is constructive to reflect on the key assumptions underlying our approach. First, we assume that cellular traits are well reflected by gene expression, which is validated by the fact that we and others (24, 88-92) can accurately classify gene expression profiles by their phenotypic labels. While transcriptional data do not directly account for posttranscriptional/translational regulation (93), they do account for downstream impacts on gene expression. Nevertheless, it is straightforward to incorporate multiomic (94-97) data to directly account for mechanisms beyond transcription. Second, we assume that transcriptional responses combine additively, which has been demonstrated to be a good approximation to control cell behavior (24). Recent work has applied VAEs to the *forward* problem of estimating nonadditive transcriptional responses to combinatorial perturbations (98), raising the possibility of going beyond the additive assumption in the future. Integration of this technique into our method to solve the *inverse* problem of mapping causes to trait phenotypes, as considered here, would require targeted experiments to train the VAE to recognize nonadditivity. Finally, we assume that our library of transcriptional responses is sufficiently large and diverse to comprehensively capture the impact of genes. Notwithstanding,

we demonstrate that enrichment analysis (99, 100) can implicate upstream transcription factors that are not included in our library.

The success of our approach has several far-reaching implications. First, it suggests that cell line perturbations in vitro are informative of the gene behavior in situ (101). Second, it shows that the genes needed to drive a phenotypic change can be distinct from those that reverse the change, a hallmark of complex systems with nonlinear dynamics. This is a consequence of our optimization model acknowledging the fundamentally different network impacts of reversing a knockdown versus overexpressing a given gene, which is consistent with persistent responses to transient perturbations observed in gene regulatory networks (102-104). Third, the success of TWAVE suggests that gene expression can be represented in a low-dimensional space (87), which might be a general feature across many complex network systems (105). Ultimately, our approach provides a tool to investigate genotype-phenotype relationships in complex traits, which is applicable across a range of organisms and traits. In humans, our approach also lays the groundwork for the design of next-generation multitarget strategies for the treatment of complex diseases.

Materials and Methods

TWAVE Architecture. We employ a conditional variational autoencoder in PyTorch, which is tailored for the analysis of genomic data and leverages class labels (baseline or variant) to impart enhanced interpretability and classification precision. The architectural blueprint consists of three neural networks: an encoder, a decoder, and a classifier.

The encoder comprises two fully connected hidden layers, each embedding 256 and 128 units, respectively, with Rectified Linear Unit (ReLU) activation functions. The input layer takes a single gene expression profile, length aligned with the number of genes, and is concatenated with pertinent class labels. This design not only captures the intricate gene expression patterns but also incorporates class-specific information for a more nuanced latent representation. The decoder component reconstructs the input RNA-Seq data through a series of ReLU-activated layers, culminating in a sigmoid activation function. This reconstruction process aims for the faithful reconstruction of the input profile from the encoded latent space. Simultaneously, the classifier, featuring a linear layer, facilitates class predictions grounded in the extracted latent representation.

TWAVE Training and Sampling. During the training phase, a set of loss functions drives the optimization process. The combination of reconstruction loss and Kullback–Leibler divergence loss is deployed, ensuring a balance between accurate data reproduction and the regularization of the latent space. The training regimen spans $500 - 10^4$ epochs, depending on the dataset, with minibatches consisting of 50 to 200 samples. We use an Adam optimizer, with a learning rate set at 0.0001.

For sampling, we generate synthetic profiles within the latent space of TWAVE. Latent vectors are obtained by sampling from clusters corresponding to distinct class labels. The latent vectors corresponding to different classes are then clustered, and marginal distributions of baseline and variant profiles in the latent space are extracted using kernel density estimation (KDE). The bandwidth parameter for our KDE is set to 0.2 to control the smoothness of the estimated density. We then sample new latent space points from these two clusters using our density estimator, and these points are decoded to the full gene expression space with the decoder, producing synthetic profiles for both class labels.

Bayesian Inference of Causal Eigengenes. We perform Bayesian inference of causal eigengenes from the gene expression data projected onto the top d = 200 eigengenes, $X = YU_{n \times d}$, where we observe that these eigengenes account for over the overwhelming majority of the variance in all traits. In *SIAppendix*, Fig. S1, we show that the remaining percent variance after d = 200 eigengenes is less than 1% for all traits except (lung) cancer metastasis and type 1 diabetes, where the remaining eigengenes account for about 22% and 23.5%

of the total variance, respectively. Our choice *d* trims the number of eigengenes to a tractable one that allows for relatively fast Monte-Carlo optimization.

Our inference procedure adapts the fine-mapping of causal variants in GWAS (17) to eigengenes. The first step of the fine-mapping is a logistic regression using class labels. We fit the log-odds ratios ζ to the data X with effect sizes β :

$$\zeta = \log \frac{\rho}{1 - \rho} = \chi \beta + \epsilon, \ \epsilon \sim \mathcal{N}(0, \sigma^2),$$
[8]

where ϵ is a noise vector, the vector ρ denotes the probabilities that each (binary) label is 1, and the log function is taken element-wise. In this scheme, we seek to infer an optimal *d*-dimensional vector of causal effects $\gamma = (\gamma_i), \gamma_i \in \{0, 1\}$, which take on values of zero or one depending on whether eigengene *i* is causal or not.

It can be shown that the likelihood of ζ , X given γ is

$$p(\zeta, X|\gamma) = \mathcal{N}(\xi|0, R + R\Sigma_{\gamma}R), \qquad [9]$$

where $\xi = X^T \zeta / d$ is the z-value, $R = X^T X / \sqrt{d\sigma^2}$ is the eigengene correlation matrix, $\Sigma_{\gamma} = ds^2 \operatorname{diag}(\gamma)$, and *s* is a hyperparameter. We set an initial s = 0.05 as in FINEMAP (17). Taking a uniform prior on the number of causal effects *k*,

$$q_k = (1/d)^k (1 - 1/d)^{d-k}$$
, [10]

we can then express the posterior distribution of causal effects given our data as

$$p(\gamma|\zeta, X) = q_k p(\zeta, X|\gamma).$$
 [11]

We use MCMC to optimize this distribution over *s* and γ , though other techniques such as FINEMAP or the sum of single effects (SuSiE) model could be employed as well (19). Posterior inclusion probabilities are calculated as an average over MCMC samples of the posterior distribution

$$PIP_{i} = p(\gamma_{i}|\zeta, X) = \frac{1}{N} \sum_{j=1}^{N} \gamma_{i}^{(j)},$$
[12]

where *N* is the number of samples. The Monte Carlo steps consist of flipping a causal effect (eigengene) on or off at random and we attempt 10^5 steps with a burn-in period of 10^3 steps, according to a Metropolis acceptance criterion.

Maximum Entropy Null Graph Model. As a null model for our gene concurrence graph, we construct a maximum entropy graph constrained by the row and column sums of our matrix *A*. The null model *G* maximizes the entropy $S(G) = -\sum_{G} P(G) \ln P(G)$, where P(G) is the canonical distribution $P(G) \propto \exp[-\sum_{i} \beta_{i}k_{i} - \sum_{i} \gamma_{i}\kappa_{i}]$. Here, $k_{i} = \sum_{j} A_{ij}$ is the row sum and $\kappa_{i} = \sum_{j} A_{ji}$ is the column sum of node *i*, whereas β_{i} and γ_{i} are the respective Lagrange multipliers that enforce k_{i} and κ_{i} to be fixed as all other degrees of freedom equilibrate. The row and column sums sequences must follow the maximum entropy conditions

$$k_i = \sum_{j \neq i} \frac{e^{\beta_i + \gamma_j}}{1 + e^{\beta_i + \gamma_j}}, \quad \kappa_j = \sum_{i \neq j} \frac{e^{\beta_j + \gamma_i}}{1 + e^{\beta_j + \gamma_i}}.$$
 [13]

- K. A. Frazer, S. S. Murray, N. J. Schork, E. J. Topol, Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.* **10**, 241–251 (2009).
- 2. E. A. Boyle, Y. I. Li, J. K. Pritchard, An expanded view of complex traits: From polygenic to
- omnigenic. *Cell* **169**, 1177–1186 (2017). 3. K. Watanabe *et al.*, A global overview of pleiotropy and genetic architecture in complex traits.
- Nat. Genet. 51, 1339-1348 (2019).
 O. Canela-Xandri, K. Rawlik, A. Tenesa, An atlas of genetic associations in UK Biobank. Nat. Genet. 50, 1593-1599 (2018).
- J. Z. Liu *et al.*, Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* 47, 979–986 (2015).
- E. E. Laing *et al.*, Blood transcriptome based biomarkers for human circadian phase. *eLife* 6,
- e20214 (2017). 7. J. Wang *et al.*, CAUSALdb: A database for disease/trait causal variants identified using summary
- J. Wang et al., CAOSACOD: A database for disease/trait causal variants identified using summary statistics of genome-wide association studies. *Nucleic Acids Res.* 48, D807–D816 (2020).
- B. Pasaniuc, A. L. Price, Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.* 18, 117–127 (2017).

These conditions can be solved for the Lagrange multipliers iteratively as

$$\beta_{i}^{(\ell+1)} = \log k_{i} - \log \sum_{j \neq i} r(\beta_{i}^{(\ell)}, \gamma_{j}^{(\ell)})$$
[14]

and

$$\gamma_j^{(\ell+1)} = \log \kappa_j - \log \sum_{i \neq j} r(\beta_j^{(\ell)}, \gamma_i^{(\ell)}), \quad [15]$$

where ℓ denotes the iteration and $r(x, y) = 1/(e^{-y} + e^x)$. From here, the link probabilities in Eq. **7** can be computed as

$$p_{ij} = rac{1}{1 + e^{-(eta_i + \gamma_j)}}.$$
 [16]

Complex Disease Data Curation. We obtained our seven datasets from GEO (Table 1). The expression matrices, originally in raw counts, are curated keeping those genes and samples that meet the following criterion: average counts in a gene >5 and total counts in a sample $>10^5$. The data were normalized to the number of transcripts per million (N_{TPM}) using reference transcript lengths mapped from the ENSEMBL gene database, and the final expression data were saved as $\log_{10}(N_{TPM} + 10^{-10}) + 10$. Labels, whether they are baseline or variant, were one-hot encoded.

Transcriptional Response Library. The transcriptional response data was curated as described in ref. 24. The list of GEO series accession numbers and associated gene perturbations are listed in Dataset S1. The inclusion of a knockdown in the library does not imply the inclusion of its overexpression and vice versa.

Data, Materials, and Software Availability. Raw gene expression counts data are available through GEO. Relevant accession numbers are included in Table 1 and Dataset S1. The software and processed data for employing the method are available from the GitHub repository (106). Source data for training TWAVE are stored on Dryad (107).

ACKNOWLEDGMENTS. This work was supported by NIH/NCI grant No. P50-CA221747 through the Malnati Brain Tumor Institute and leveraged research from NSF grant No. MCB-2206974. We also acknowledge support from the NSF-Simons National Institute for Theory and Mathematics in Biology (NSF Grant No. DMS-2235451 and Simons Foundation Grant No. MP-TMPS-00005320) and the use of Quest High-Performance Computing Cluster at Northwestern University.

Author affiliations: ^aDepartment of Physics and Astronomy, Northwestern University, Evanston, IL 60208; ^bCenter for Network Dynamics, Northwestern University, Evanston, IL 60208; ^cDepartment of Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, IL 60208; ^d Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL 60208; ^eNational Institute for Theory and Mathematics in Biology, Chicago, IL 60611; and ^fChemistry of Life Processes Institute, Northwestern University, Evanston, IL 60208

- A. M. Glazier, J. H. Nadeau, T. J. Aitman, Finding genes that underlie complex traits. Science 298, 2345–2349 (2002).
- S. Tang et al., Novel Variance-Component TWAS method for studying complex human diseases with applications to Alzheimer's dementia. PLoS Genet. 17, e1009482 (2021).
- 11. P. Visscher *et al.*, 10 years of GWAS discovery: Biology, function, and translation. *Am. J. Hum. Genet* **101**, 5–22 (2017).
- 12. E. Uffelmann et al., Genome-wide association studies. Nat. Rev. Methods Primers. 1, 59 (2021).
- V. Tam et al., Benefits and limitations of genome-wide association studies. Nat. Rev. Genet. 20, 467–484 (2019).
- E. Gamazon et al., A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. 47, 1091–1098 (2015).
- M. Wainberg et al., Opportunities and challenges for transcriptome-wide association studies. Nat. Genet. 51, 592–599 (2019).
- W. Wei, G. Hemani, C. Haley, Detecting epistasis in human complex traits. *Nat. Rev. Genet.* 15, 722–733 (2014).

- C. Benner *et al.*, FINEMAP: Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* 32, 1493–1501 (2016).
- Y. Zou, P. Carbonetto, G. Wang, M. Stephens, Fine-mapping from summary data with the "Sum of Single Effects" model. *PLoS Genet.* 18, e1010299 (2022).
- G. Wang, A. Sarkar, P. Carbonetto, M. Stephens, A simple new approach to variable selection in regression, with application to genetic fine mapping. J. R. Stat. Soc. Ser. B Stat. Methodol. 82, 1273–1300 (2020).
- A. Hutchinson, J. Asimit, C. Wallace, Fine-mapping genetic associations. Hum. Mol. Genet. 29, R81-R88 (2020).
- K. K. H. Farh *et al.*, Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
- A. R. Sonawane, S. T. Weiss, K. Glass, A. Sharma, Network medicine in the age of biomedical big data. Front. Genet. 10, 294 (2019).
- G. Zhang, N. M. Roberto, D. Lee, S. R. Hahnel, E. C. Andersen, The impact of species-wide gene expression variation on *Caenorhabditis elegans* complex traits. *Nat. Commun.* 13, 3462 (2022).
- T. P. Wytock, A. E. Motter, Cell reprogramming design by transfer learning of functional transcriptional networks. *Proc. Natl. Acad. Sci. U.S.A.* 121, e2312942121 (2024).
- O. Alter, P. O. Brown, D. Botstein, Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U.S.A.* 97, 10101-10106 (2000).
- S. Kauffman, Metabolic stability and epigenesis in randomly constructed genetic nets. J. Theoret. Biol. 22, 437–467 (1969).
 S. Luczer, G. Eicher, V. D. Kart, D. F. Lacher, Call Schwarz, Eich dimensional attractionation of several stability of the several stability of th
- S. Huang, G. Eichler, Y. Bar-Yam, D. E. Ingber, Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Phys. Rev. Lett.* 94, 128701 (2005).
- S. Huang, S. Kauffman, "Complex gene regulatory networks from structure to biological observables: Cell fate determination" in *Encyclopedia of Complexity and Systems Science*, R. Meyers, Ed. (Springer, New York, NY, 2009), pp. 1180–1213.
- R. Shemirani, S. Wenric, E. Kenny, J. L. Ambite, EPS: Automated feature selection in case-control studies using extreme pseudo-sampling. *Bioinform.* 37, 3372–3373 (2021).
- S. Wenric, R. Shemirani, Using supervised learning methods for gene selection in RNA-Seq case-control studies. Front. Genet. 9, 297 (2018).
- T. A. Karginov, A. Ménoret, A. T. Vella, Optimal CD8+ T cell effector function requires costimulation-induced RNA-binding proteins that reprogram the transcript isoform landscape *Nat. Commun.* 13, 3540 (2022).
- S. Yu, V. N. Kim, A tale of non-canonical tails: Gene regulation by post-transcriptional RNA tailing. Nat. Rev. Mol. Cell Biol. 21, 542-556 (2020).
- X. Feng et al., Integrative analysis of the expression profiles of whole coding and non-coding RNA transcriptomes and construction of the competing endogenous RNA networks for chronic obstructive pulmonary disease. Front. Genet. 14, 1050783 (2023).
- E. M. Mushaben, G. K. Hershey, M. W. Pauciulo, W. C. Nichols, T. D. Le Cras, Chronic allergic inflammation causes vascular remodeling and pulmonary hypertension in BMPR2 hypomorph and wild-type mice. *PLoS One* 7, e32468 (2012).
- S. H. Park *et al.*, Modification of hemodynamic and immune responses to exposure with a weak antigen by the expression of a hypomorphic BMPR2 gene. *PloS One* 8, e55180 (2013).
- Y. Zhu, W. Wang, X. Wang, Roles of transcriptional factor 7 in production of inflammatory factors for lung diseases. J. Transl. Med. 13, 273 (2015).
- N. Hernandez-Pacheco et al., Genome-wide association study of inhaled corticosteroid response in admixed children with asthma. Clin. Exp. Allergy 49, 789-798 (2019).
- A. K. Kheirallah, C. H. de Moor, A. Faiz, I. Sayers, I. P. Hall, Lung function associated gene Integrator Complex subunit 12 regulates protein synthesis pathways. *BMC Genom.* 18, 1–20 (2017).
- X. Li, S. Ye, Y. Lu, Long non-coding RNA NEAT1 overexpression associates with increased exacerbation risk, severity, and inflammation, as well as decreased lung function through the interaction with microRNA-124 in asthma. J. Clin. Lab. Anal. 34, e23023 (2020).
- D. G. Roy et al., Methionine metabolism shapes T helper cell responses through regulation of epigenetic reprogramming. Cell Metab. 31, 250–266 (2020).
- M. Shang et al., The folate cycle enzyme MTHFD2 induces cancer immune evasion through PD-L1 up-regulation. Nat. Commun. 12, 1940 (2021).
- R. D. Britt Jr. et al., Sterols and immune mechanisms in asthma. J. Allergy Clin. Immunol. 151, 47–59 (2023).
- A. García Del Río *et al.*, The mitochondrial isoform of FASTK modulates nonopsonic phagocytosis of bacteria by macrophages via regulation of respiratory complex I. *J. Immunol.* 201, 2977–2985 (2018).
- M. R. Hadjicharalambous, M. A. Lindsay, Long non-coding RNAs and the innate immune response. *Noncoding RNA* 5, 34 (2019).
- Z. Li et al., The long noncoding RNA THRIL regulates TNFα expression through its interaction with hnRNPL. Proc. Natl. Acad. Sci. U.S.A. 111, 1002–1007 (2014).
- Z. Su, D. Huang, Alternative splicing of pre-mRNA in the control of immune activity. *Genes* 12, 574 (2021).
- 47. P. Zeng *et al.*, Statistical analysis for genome-wide association study. *J. Biomed. Res.* **29**, 285 (2015).
- S. Chatterjee, P. Diaconis, A. Sly, Random graphs with a given degree sequence. Ann. Appl. Probab. 21, 1400–1435 (2011).
- 49. G. Bianconi, Entropy of network ensembles. Phys. Rev. E 79, 036114 (2009).
- K. M. Magnaye et al., A-to-I editing of miR-200b-3p in airway cells is associated with moderate-tosevere asthma. Eur. Respir. J. 58, 2003862 (2021).
- V. Tubita *et al.*, Role of microRNAs in inflammatory upper airway diseases. *Allergy* **76**, 1967–1980 (2020).
- T. Xia, J. Ma, Y. Sun, Y. Sun, Androgen receptor suppresses inflammatory response of airway epithelial cells in allergic asthma through MAPK1 and MAPK14. *Hum. Exp. Toxicol.* 41, 09603271221121320 (2022).
- C. R. Keenan *et al.*, Polycomb repressive complex 2 is a critical mediator of allergic inflammation. *JCl Insight* 4, e127745 (2029).
- S. N. Georas, P. Donohue, M. Connolly, M. E. Wechsler, JAK inhibitors for asthma. J. Allergy. Clin. Immunol. 148, 953–963 (2021).
- L. Ye et al., A critical role for c-Myc in group 2 innate lymphoid cell activation. Allergy 75, 841–852 (2020).

- V. Poli et al., MYC-driven epigenetic reprogramming favors the onset of tumorigenesis by inducing a stem cell-like state. Nat. Commun. 9, 1024 (2018).
- S. Patange *et al.*, MYC amplifies gene expression through global changes in transcription factor dynamics. *Cell Rep.* 38, 110292 (2022).
- C. Y. Wang et al., Induced pluripotent stem cells without c-Myc reduce airway responsiveness and allergic reaction in sensitized mice. *Transplant.* 96, 958–965 (2013).
- P. Borger *et al.*, Impaired translation of CCAAT/enhancer binding protein α mRNA in bronchial smooth muscle cells of asthmatic patients. *JACI* **123**, 639–645 (2009).
- S. Chung *et al.*, FoxO1 regulates allergic asthmatic inflammation through regulating polarization of the macrophage inflammatory phenotype. *Oncotarget* 7, 17532 (2016).
- X. Rao et al., MIR-493-5p inhibits Th9 cell differentiation in allergic asthma by targeting FOXO1. Respir. Res. 23, 286 (2022).
- F. J. Ren, X. Y. Cai, Y. Ao, G. Y. Fang, JunB: A paradigm for Jun family in immune response and cancer. Front. Cell. Infect. Microbiol. 13, 1222265 (2023).
- S. Rojo-Tolosa et al., Influence of genetics on the response to omalizumab in patients with severe uncontrolled asthma with an allergic phenotype. Int. J. Mol. Sci. 24, 7029 (2023).
- C. Meng *et al.*, Ten-eleven translocation 2 modulates allergic inflammation by 5hydroxymethylcytosine remodeling of immunologic pathways. *Hum. Mol. Genet.* **30**, 1985–1995 (2021).
- H. Zhang et al., Targeting epithelial cell-derived TWIST1 alleviates allergic asthma. Cell. Signal. 102, 110552 (2023).
- 66. A. Tsherniak *et al.*, Defining a cancer dependency map. *Cell* **170**, 564–576 (2017).
- P. S. Dischinger et al., NF1 deficiency correlates with estrogen receptor signaling and diminished survival in breast cancer. NPJ Breast Cancer 4, 29 (2018).
- J. Hu et al., Sox5 contributes to prostate cancer metastasis and is a master regulator of TGF-βinduced epithelial mesenchymal transition through controlling Twist1 expression. Br. J. Cancer 118, 88–97 (2018).
- N. Malik *et al.*, The transcription factor CBFB suppresses breast cancer through orchestrating translation and transcription. *Nat. Commun.* **10**, 2071 (2019).
- Y. L. Chen et al., LnCRNA SLCO4A1-AS1 suppresses lung cancer progression by sequestering the TOX4-NTSR1 signaling axis. J. Biomed. Sci. 30, 80 (2023).
 T. Elicit A. Scrait M. Sci Ladetaine M. Minde Langering for DD044. In Action 4. Science M. Science
- T. Elsir, A. Smits, M. S. Lindström, M. Nistér, Transcription factor PROX1: Its role in development and cancer. *Cancer Metastasis Rev.* 31, 793–805 (2012).
- Z. Cheng et al., Knockdown of EHF inhibited the proliferation, invasion and tumorigenesis of ovarian cancer cells. *Mol. Carcinog.* 55, 1048–1059 (2016).
- M. Z. Kamran, P. Patil, R. P. Gude, Role of STAT3 in cancer metastasis and translational advances. Biomed Res. Int. 2013, 421821 (2013).
- L. Zhao et al., CTCF promotes epithelial ovarian cancer metastasis by broadly controlling the expression of metastasis-associated genes. Oncotarget 8, 62217 (2017).
- K. Takeda et al., Targeted disruption of the mouse Stat3 gene leads to early embryonic lethality. Proc. Natl. Acad. Sci. U.S.A. 94, 3801–3804 (1997).
- A. M. Fedoriw, P. Stein, P. Svoboda, R. M. Schultz, M. S. Bartolomei, Transgenic RNAi reveals essential function for CTCF in H19 gene imprinting. *Science* 303, 238–240 (2004).
- W. Chen, X. Zhang, K. Birsoy, R. G. Roeder, A muscle-specific knockout implicates nuclear receptor coactivator MED1 in the regulation of glucose and energy metabolism. *Proc. Natl. Acad. Sci. U.S.A.* 107, 10196–10201 (2010).
- D. Liu et al., Evaluation of the oxidative stress-related genes ALOX5, ALOX5AP, GPX1, GPX3 and MPO for contribution to the risk of type 2 diabetes mellitus in the Han Chinese population. Diabetes Vasc. Dis. Res. 15, 336–339 (2018).
- M. M. Heemskerk et al., Increased PUFA content and 5-lipoxygenase pathway expression are associated with subcutaneous adipose tissue inflammation in obese women with type 2 diabetes. Nutrients 7, 7676-7690 (2015).
- C. Argmann *et al.*, Biopsy and blood-based molecular biomarker of inflammation in IBD. *Gut* 72, 1271–1287 (2023).
- A. Samal, S. Jain, The regulatory network of *E. coli* metabolism as a Boolean dynamical system exhibits both homeostasis and flexibility of response. *BMC Syst. Biol.* 2, 1–18 (2008).
- R. S. Wang, A. Saadatpour, R. Albert, Boolean modeling in systems biology: An overview of methodology and applications. *Phys. Biol.* 9, 055001 (2012).
- H. Sizek, A. Hamel, D. Deritei, S. Campbell, E. Ravasz Regan, Boolean model of growth signaling, cell cycle and apoptosis predicts the molecular mechanism of aberrant cell cycle progression driven by hyperactive PI3K. *PLoS Comput. Biol.* **15**, e1006402 (2019).
- J. C. Rozum, J. Gómez Tejeda Zañudo, X. Gan, D. Deritei, R. Albert, Parity and time reversal elucidate both decision-making in empirical models and attractor scaling in critical Boolean networks. Sci. Adv. 7, eabf8124 (2021).
- E. Newby, J. G. Tejeda Zañu, R. Albert, Structure-based approach to identifying small sets of driver nodes in biological networks. *Chaos* 32, 063102 (2022).
- E. Sullivan et al., Boolean modeling of mechanosensitive epithelial to mesenchymal transition and its reversal. *iScience* 26, 106321 (2023).
- 87. K. Hari et al., Low dimensionality of phenotypic space as an emergent property of coordinated teams in biological regulatory networks. *iScience* 28, 111730 (2025).
- K. F. Mahin *et al.*, PanClassif: Improving pan cancer classification of single cell RNA-Seq gene expression data using machine learning. *Genomics* **114**, 110264 (2022).
- T. Lee, H. Lee, Prediction of Alzheimer's disease using blood gene expression data. Sci. Rep. 10, 3485 (2020).
- B. Kegerreis et al., Machine learning approaches to predict lupus disease activity from gene expression data. Sci. Rep. 9, 9617 (2019).
- 91. F. Alharbi, A. Vakanski, Machine learning methods for cancer classification using gene expression data: A review. *Bioengineering* **10**, 173 (2023).
- W. Cookson, L. Liang, G. Abecasis, M. Moffatt, M. Lathrop, Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.* 10, 184–194 (2009).
- D. Zhang et al., Global and gene-specific translational regulation in Escherichia coli across different conditions. PLoS Comput. Biol. 18, e1010641 (2022).
- S. Albaradei *et al.*, MetaCancer: A deep learning-based pan-cancer metastasis prediction model developed using multi-omics data. *Comput. Struct. Biotechnol. J.* 19, 4404–4411 (2021).

- A. Ebrahim et al., Multi-omic data integration enables discovery of hidden biological regularities. Nat. Commun. 7, 13091 (2016).
- A. Gayoso et al., Joint probabilistic modeling of single-cell multi-omic data with totalVI. Nat. Methods 18, 272-282 (2021).
- Y. Lu, M. Oliva, B. L. Pierce, J. Liu, L. S. Chen, Integrative cross-omics and cross-context analysis elucidates molecular links underlying genetic effects on complex traits. *Nat. Commun.* 15, 2383 (2024).
- M. Lotfollahi *et al.*, Predicting cellular responses to complex perturbations in high-throughput screens. *Mol. Syst. Biol.* **19**, e11517 (2023).
- 99. Z. Xie, Gene set knowledge discovery with enrichr. Curr. Protoc. 1, e90 (2021).
- K. Glass, M. Girvan, Finding new order in biological functions from the network structure of gene annotations. *PLoS Comput. Biol.* **11**, e1004565 (2015).
- M. Lotfollahi, F. Wolf, F. Theis, scGen predicts single-cell perturbation responses. *Nat. Methods.* 16, 715–721 (2019).

- Y. Zhao, T. P. Wytock, K. A. Reynolds, A. E. Motter, Irreversibility in bacterial regulatory networks. Sci. Adv. 10, eado3232 (2024).
- R. Zhu, J. M. del Rio-Salgado, J. Garcia-Ojalvo, M. B. Elowitz, Synthetic multistability in mammalian cells. *Science* 375, eabg9765 (2022).
- K. Joshi et al., Non-Markovian memory in a bacterium. bioRxiv [Preprint] (2023). https://doi.org/ 10.1101/2023.05.27.542601 (Accessed 13 August 2024).
- V. Thibeault, A. Allard, P. Desrosiers, The low-rank hypothesis of complex systems. Nat. Phys. 20, 294–302 (2024).
- B. Kuznets-Speck, B. K. Ogonor, T. P. Wytock, A. E. Motter, Code for "Generative prediction of causal gene sets responsible for complex traits". GitHub. https://github.com/biophysben/ Generative-prediction-of-causal-gene-sets-responsible-for-complex-traits. Deposited 1 April 2025.
- B. Kuznets-Speck, B. K. Ogonor, T. P. Wytock, A. E. Motter, Data for "Generative prediction of causal gene sets responsible for complex traits". Dryad. https://doi.org/10.5061/dryad.s4mw6m9hf. Deposited 1 April 2025.